Rochester Institute of Technology
RIT Scholar Works

Theses | Thesis/Dissertation Collections

2-20-2012

Development of a process for identification of the operational mode of industrial sites using high dimensional multi-modal data

Jake Clements

Follow this and additional works at: http://scholarworks.rit.edu/theses

Recommended Citation
Clements, Jake, "Development of a process for identification of the operational mode of industrial sites using high dimensional multi-modal data" (2012). Thesis. Rochester Institute of Technology. Accessed from

This Dissertation is brought to you for free and open access by the Thesis/Dissertation Collections at RIT Scholar Works. It has been accepted for inclusion in Theses by an authorized administrator of RIT Scholar Works. For more information, please contact ritscholarworks@rit.edu.

Development of a Process for Identification of the Operational Mode of Industrial Sites Using High Dimensional Multi-modal Data

by

Jake Clements

B.A. Computer Science, SUNY Geneseo, 2002

A dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the Chester F. Carlson Center for Imaging Science
Rochester Institute of Technology

February 20, 2012

Signature of the Author _____

Accepted by _____
Coordinator, Ph.D. Degree Program Date

CHESTER F. CARLSON CENTER FOR IMAGING SCIENCE
ROCHESTER INSTITUTE OF TECHNOLOGY
ROCHESTER, NEW YORK

CERTIFICATE OF APPROVAL

Ph.D. DEGREE DISSERTATION

The Ph.D. Degree Dissertation of Jake Clements
has been examined and approved by the
dissertation committee as satisfactory for the
dissertation required for the
Ph.D. degree in Imaging Science

_____
Dr. John Schott, Dissertation Advisor

_____
Dr. Peter Bajorski

_____
Dr. Emmett Ientilucci

_____
Dr. Matthew Coppenbarger

_____
Date

DISSERTATION RELEASE PERMISSION
ROCHESTER INSTITUTE OF TECHNOLOGY
CHESTER F. CARLSON CENTER FOR IMAGING SCIENCE

Title of Dissertation:

Development of a Process for Identification of the Operational Mode of Industrial Sites Using High Dimensional Multi-modal Data

I, Jake Clements, hereby grant permission to Wallace Memorial Library of R.I.T. to reproduce my thesis in whole or in part. Any reproduction will not be for commercial use or profit.

Signature _____

Date

Development of a Process for Identification of the Operational Mode of Industrial Sites Using High Dimensional Multi-modal Data

by

Jake Clements

Submitted to the
Chester F. Carlson Center for Imaging Science
in partial fulfillment of the requirements
for the Doctor of Philosophy Degree
at the Rochester Institute of Technology

Abstract

Many algorithms exist to determine the physical contents of an image. Target detection or anomaly detection algorithms, for example, use statistical and geometric approaches in high dimensional space to locate objects within a scene. Instead of target detection, however, it has become of interest of late to delve deeper into the field of remote sensing in order to perform *process detection*. Process detection refers to the ability to identify the operational mode of an industrial facility. To accurately complete this task will require a new set of analysis tools.

This thesis discusses a method that can be used to perform process detection with multi-modal remotely sensed data. Using a local industrial facility, operational modes were identified, as well as the subtle differences between them. Combinations of hourly data, sparse data, and latent variables were combined through analytical tools and a prediction of the process taking place at different moments was performing using both real and simulated data sets.

An advanced analyst environment is also discussed, with a few demonstrations from a test environment developed by a small team at RIT. Temporal analysis, multi-modal data integration, and the use of process models to make latent observables are discussed. This thesis shows the utility of such an environment and demonstrates the need for the further development.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In the remote sensing community image analysts are tasked with finding out what information is contained within an image. Typically this is limited to what physical objects are in the image, where they are, and how many of them are there. Through these studies one can often tell the difference between a college and a high school, or a nuclear plant and a water treatment plant. Clearly it is of interest to the intelligence community to take this information to the next level. Now that one can say with certainty that one is looking at a nuclear plant, it is desirable to be able to determine whether the plant is reprocessing spent fuel rods or enriching uranium to weapons grade. Such differences will manifest themselves through several types of signals, and it is important to have a quantitative method that effectively combines the information from those signals into a probability of each possible state.

Shortly after September 11, 2001 our government tried to convince the country to support it in its invasion of Iraq. While full support was never obtained, the government was able to obtain enough of a backing by showing Congress evidence of weapons of mass destruction in Iraq. The invasion followed, the war raged on for several years, and none of the weapons were ever found. This is a situation in which our intelligence completely failed and it demonstrates the need for new approaches to intelligence analysis.

Presently there is so much information being placed in front of the image analysts

that it has become difficult for them to sort it all out and ably separate the relevant information from the irrelevant. This thesis discusses an approach that has been developed, as well as a hypothetical computer environment, that can assist analysts so that they can more easily take in information, sift through historical data, and draw conclusions with higher confidence. Through the work described in this thesis these findings can now be linked together with other sources of information so that an analyst can accurately describe the processes taking place within a facility of interest. This approach is relatively simple in terms of its technical interface to accommodate the broad expertise range (often non technical) of most analysts.

Objectives

The AANEE Project

One can imagine a data repository in which all forms of intelligence data are stored and geo-referenced, an environment in which to seamlessly interact with data, and advanced algorithms to assist the analyst in various tasks such as target detection or site identification. That ideal environment is still a long way off, but it has been labeled part of the Advanced ANalyst Exploitation Environment (AANEE) project. This project has been split into three portions: development of a three dimensional interactive environment in which to store the data, advanced registration techniques to relate the data, and development of a process for testing and evaluation of exploitation algorithms that utilize the other two as well as feed in new data. The first two portions of the project are being undertaken by other researchers but are mentioned here because they are intimately related to the work being proposed herein as shown in Figure 1.1.

The ideal AANEE environment is one in which there are 3D models of all buildings and terrain. A large variety of data types would be available, including all forms of imagery, human intelligence, and other signal data such as RF and seismic. An analyst should be able to listen to a building "talk" about its purpose and history while analyzing the different forms of information available pertaining to it. It

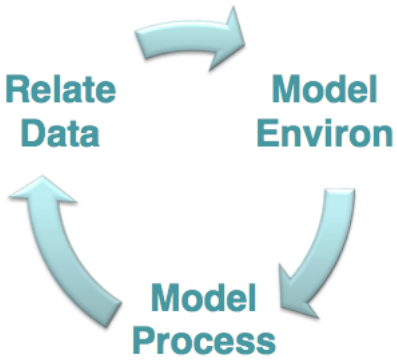


Figure 1.1: *A three pronged approach to interactive site modeling with each piece being highly interrelated to the other two.*

should be possible to scroll through time, to see changes that have been made either in the structure or in the signals present. Any combination of data should then be able to be extracted for more in depth analysis through advanced algorithms that are unable to run in real time. The output from these algorithms, however, should then be able to be fed back in to the environment, so as to provide more data to a future analyst.

Such an environment would need a user interface that is easily extensible to many disciplines. For each building to know its function it will be necessary to have process models of the facility. Developing process models for all aspects of a facility will require a subject matter expert. In order to bring in large quantities of data it will be necessary to have a data reader for each datatype. Lastly this should not be limited to a single facility, but should instead cover the world, so that all sites can be monitored, either to track weapon production at a foreign site or to track the all hazardous chemical usage at domestic sites.

The ideal scenario is too massive for a university research group to undertake. A fully functioning version of this idea is something this thesis seeks to convince others to develop by exploring some of the ways to bring data together and the use of analysis tools that can help. Some parts of the environment have been implemented by the AANEE team using a mixture of real and simulated data to explore potential

tools envisioned for this dream. Since one of the goals of the AANEE project is to have the process described in this thesis implemented in the interactive environment, some of the preliminary tests will be demonstrated throughout this thesis.

A Method for Process Identification

The primary objective of the proposed research is to develop and demonstrate a process for the in depth analysis of multi-source, multi-angle, multi-temporal data of a single site. This method starts with the assumption that three dimensional data registration exists and that there is already an interactive environment in place in which to examine the data. The process then follows a series of steps, beginning with intimately identifying a target site, followed by signal analysis and quantitative analysis of the signals.

Site Identification

When an analyst is first tasked with a site he/she needs to follow a series of steps to learn everything there is to know. Site identification refers to the manner in which an analyst can determine the various modalities of a facility. This is broken down into three parts. First, infrastructure analysis is done so that the various subprocesses can be identified. These are then combined together to reveal the main function as well as the flow of all of the materials as they traverse the site. The different operational states can then be described (production level, on/off, efficiency - all site dependent), so that differences among them become evident.

Signal Analysis

It is desirable to find a way to integrate data driven techniques (such as target detection from imagery) with process models derived from domain experts. Analysts should be able to link these through logical processes that are designed to identify, model, and/or predict observable signals. Signal analysis is a preparation technique that will enable an analyst to know which remote sensing modalities can be used for

a particular site. The observable signals across a broad range of modalities (types of intelligence data or INTs) that may allow one to differentiate between modes are identified. One will have to determine the relationship each observable signal has to the operational modes of the site. These signals are a representation of the relationship of the different processes and their different manifestations. The range of values each signal may have and how these are manifested in the different types of sensor modalities needs to be understood, as well as how a change in one signal may affect other signals.

Analysis Tools

The final stage of this process is combining the observable signals in some form of analysis technique to predict the operational mode of the site. For this project three different algorithms, each building off the relational statistics of the various observable signals in different complexity, are examined. The first and most basic is template matching, which will be an easy to build model requiring relatively small amounts of data, as long as the physical relationships are known or approximated. Second is a geometrical approach to clustering vector representations of the data, which requires significant amounts of data collected over weeks or months. Lastly direct statistical methods are discussed which can make the most precise calculations, but may require months or years of data collection. Since waiting months is not desirable (even weeks can be too long), it may become necessary to utilize a cooperative site. This will enable one to use simulated, estimated, or statistical data in order to determine the relationship each observable has with each mode with a fairly high initial degree of confidence.

In regards to confidence, one may want to incorporate a confidence quantity in each observable. This practice will be explored as well. Many different factors may be taken into consideration when talking about confidence. For example, data collected from a brand new experimental sensor may not be as reliable as data collected from a well researched, well calibrated sensor that has been around for years. The collected data may be sparse, ill-quantified, or simulated. Furthermore, the source

of the data (e.g., US, France, Britain) might be of concern. Calculating confidence to an exact number is not often possible as it is typically considered a quality, not a quantity. However, estimates can be made and a weighting can be given to each of the observables used for process identification.

This process is analyzed through the use of a friendly industrial site using a large amount of real data, specifically the Van Lare Wastewater Treatment plant. This plant offers a wide array of signals and operations, ranging from open air processes to subterranean input and output. This site is described in great detail in Section 3.1. The operational modes are identified and detected accurately. As a final test some artificial scenarios have been developed, and simulated data based on the real physical relationships of the variables is used to test the analysis tools in extreme situations.

The contribution of this thesis is to demonstrate the possibilities of such an environment, as well as identify the issues and requirements in both the data and the analysis tools. The current intelligence analysis methods used by our government lacks the ability to accurately determine the operational mode of a facility. The process developed here does not completely solve this problem, however it greatly increases the accuracy from near zero to around 50%.

Chapter 2

Background

2.1 3D Data Registration and Analysis

Since 3D data registration is assumed to be a conquered subject for the AANEE project, this section focuses on different analyst environments and the manner in which they handle this problem. A true solution to the problem with an easily attainable environment is not presently available, but near approximations do exist.

Image registration deals with aligning two or more images so that objects in all of the images line up together (Schott, 2007). Since not all of the information being used in this project is imagery it is necessary and prudent to extend this idea to all of the data. While precise registration is not necessary for the exploitation tools as implemented for this initial study, it would allow for a program to automatically update the results of analysis tools once they have been built for a site.

Two dimensional registration links images together at a pixel level. 3D image registration extracts the geometrical information from the data in order to recover structure. An example of this process is shown in Figure 2.1 where an image derived model is registered to a LIDAR model. LIDAR is a remote sensing paradigm used to collect 3D information, while images are 2D at each band. Using advanced photogrammetry techniques it is possible to derive low quality models from imagery. When this model is registered to a high quality LIDAR model, the result is a high

quality model with spectral information (Walli, 2010).

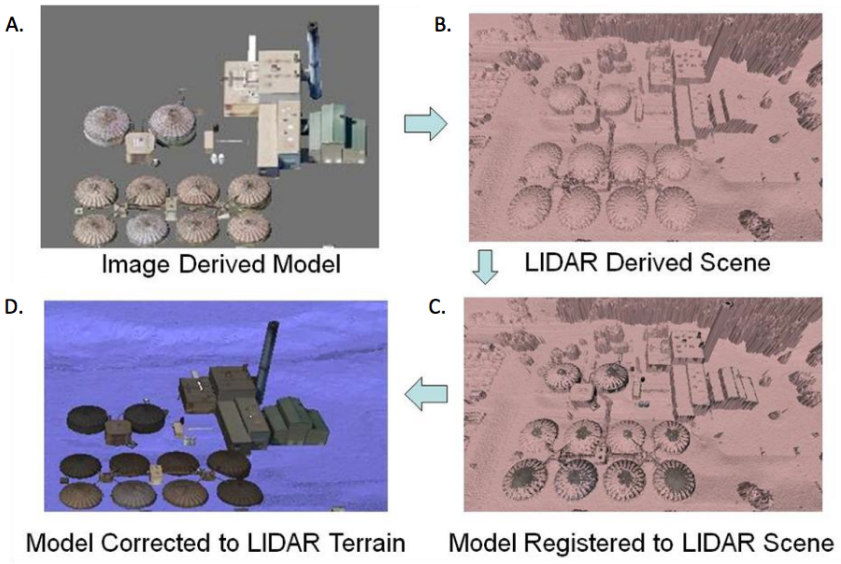


Figure 2.1: *An example of 3D registration. A is an image derived model, B is a LIDAR model. The two are shown registered to each other in C, and D is a hyrbridization of the two models: high quality 3D structure with spectral information. Images courtesy of Karl Walli.*

The goal is to find, for example, a way in which to tie a time sampled observation of a road to a newspaper article about a car accident and a thermal image of a facility. Commercial environments that attempt to do this have many deficiencies which is why construction of an environment to explore possible new tools was initiated in this thesis. For completeness, what follows in the next section is a brief discussion of some commercial analyst environments.

2.1.1 Analyst Environments

Current image analysis environments allow for some data integration, but rarely is multi-INT data brought together in the same place. There are several projects available for free download that offer image analysis tools. The University of Manchester has a project called TINA(Tina Is No Acronym), Vision Systems Group's project

is called NeatVision, which are just two examples. These programs allow for both computer vision (target detection, object recognition) and medical image analysis. The software deals mostly with grey scale images, though, with no mention of three dimensional registration which would be required for multi-modal, multi-angle data (TINA, 2008; NeatVision, 2008).

More advanced environments are available such as Google Earth and ArcGIS. These more advanced software packages allow for the importation of all sorts of image data, including three dimensional models that can be made with CAD software or extracted from LIDAR point clouds. Within these software packages are some advanced registration techniques that, while not perfect, allow analysts to view multiple data sources with only minor difficulties. Lacking from these software packages is the integration of the data with process models and/or prediction models. An analyst can point out an object and declare it to be a tank with ninety percent certainty, and the position of the object can be tracked over time, but there is nothing that relates that object to the environment in which it was found (Google Earth, 2010; ArcGIS, 2010). The remote sensing community believes that the near future will not bring about vast changes in sensor design but instead revolutionary advances in analytical environments. There is a push for sensor fusion that will allow more information to be extracted from current remote sensing systems (Gail, 2007). This effort is designed to develop and test tools that might be incorporated into such an analyst exploitation environment. It is important to be able to compare new tools based on their ease of use, performance, and how complex of a task each can handle.

Data assimilation refers to the manner in which the data is brought together and collectively analyzed. There are several tools already in existence that were used during the course of this research, the top two being ArcGIS and Google Earth. Google Earth (GoogleEarth.com) is a powerful image analysis tool that is being used by the National Geospatial Intelligence Agency (NGA) (Messinger, 2007). ArcGIS (ArcGIS.com) is a popular software package being used by several government agencies, including Monroe county officials (Lukas, 2007). Both software packages allow one to collect images of a site of interest together in one place and

layer them on top of each other. One can mark points of interest in each image and see where they line up in other images. Both were used in the early stages of this project in order to test their capabilities and limitations and to see how they compare to the final goal of the AANEE project.

ArcGIS

ArcGIS (Arc) is a well known imagery storage system that is used by several government agencies. It has similar capabilities to GE, however it does not have a free version available. The advantage that Arc and GE professional have is that they tie images together through the global positioning system (GPS). This allows for significantly more accurate importation of imagery. Some image formats have their GPS coordinates built into the image header file so that one merely has to select the image to be imported and the software takes care of the rest. This is demonstrated in Figures 2.2 and 2.3. The latest version of Arc is capable of doing much of the basic analyses required for this project, but one would still have to build and test their own analysis models and integrate them through the user interface. It is also primarily a 2D environment (Esri.com, 2010).

Figure 2.2: *Arc image of the building locations being overlaid on an RGB image. A slight misregistration can be seen, but it is the same at all angles and zooms, and easily compensated for by an analyst. Files courtesy of Monroe County.*

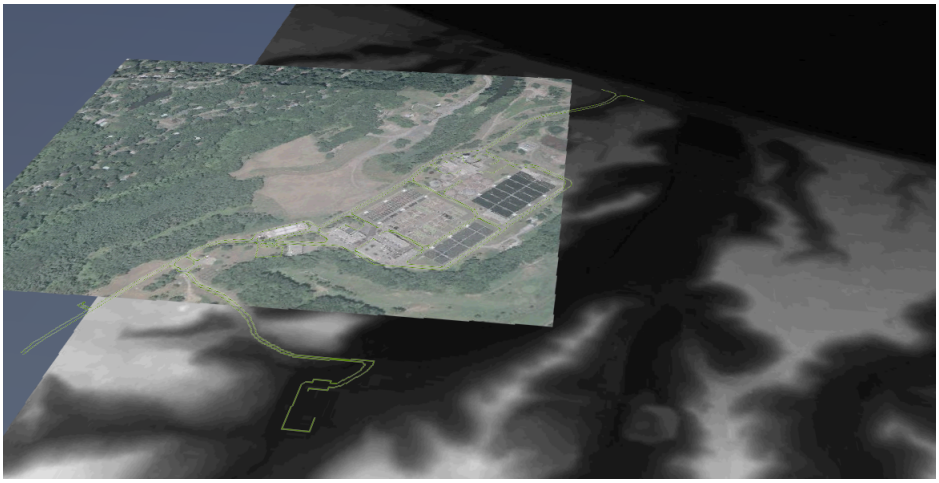


Figure 2.3: *Arc image of roads overlaid on an RGB image, placed on top of a digital elevation model. Files courtesy of Monroe County.*

Google Earth

Google Earth (GE) has both a free version and a professional version available for use. The free version, while not as powerful as the professional version, will likely provide all that one would need for this experimental research project. It provides full color, high resolution, nadir imagery of the majority of the United States and a good portion of the rest of the world. Through the image import tool one can have their own images overlaid on the Google scene as seen in Figure 2.4. Each image can be tagged with a time stamp, and points of interest can then be tracked temporally.

Because of GE's low cost, widespread use in the intelligence community, and accessibility it was decided that software developed in this research would be able to communicate with it freely. Some images and models were able to be exported from GE and imported into the exploratory AANEE software, and vice versa. This task goes outside the scope of this research, but is within the scope of the AANEE project as a whole, and was handled as a joint effort by Karl Walli and Colin Doody (Walli, 2010). The software package aided in the analysis portion of the project (the part of AANEE directly related to the research presented in this thesis) by providing a means to immerse oneself in the data, enabling the tagging of key points and the

Figure 2.4: *GE image with LWIR image laid on top of RIT campus.*

recording of observables. In order for all of this to be possible, however, an efficient manner of image storage and retrieval is necessary.

2.1.2 Image Storage

Image storage such that quick and easy retrieval is possible is a very difficult task. There are several pieces of information that are relevant to each image, such as date and time of capture, instrument used, available bands, resolution, weather, where the shot was taken from, GPS coordinates of the image corners, and the content contained within. Most database systems were originally designed to handle large quantities of alpha-numeric data. However, many are not yet ready to handle image data.

A thorough image database would need to be able to store the original input image as well as any images that are simply processed forms of the original image, and easily be able to differentiate between the two. Any tags that have been made in the image, such as points of interest or textual additions from an observer, need to be stored, as well as returned when one is trying to retrieve a different image with the same physical contents. Mehrotra suggests that processed images should be linked with the process that was used to make them (Mehrotra, 1995). While the source of this proposed system is over ten years old, only temporary methods have been suggested; a permanent solution to the problem has yet to be discovered.

Attempting to accomplish all of these tasks is a doctoral project in itself and goes beyond the scope of this project.

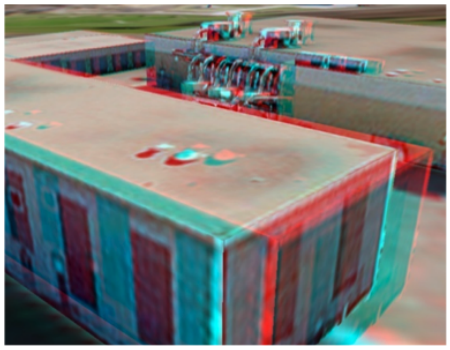
Esri and Oracle have designed a database system that is used in ArcGIS 10. It does allow for the integration of spatial data and business information. It can also handle multiuser access to a large quantity of image data. However, the system is not to a point where it can handle running image analysis algorithms on its imagery and efficiently store the results. Esri has recently partnered with the NGA to begin working on such a system (Esri.com, 2010).

2.1.3 The AANEE Environment

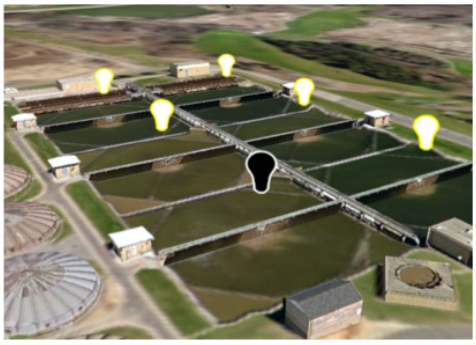
The limitations of current analyst environments demonstrate a need for the AANEE project. Using interactive gaming technology, the team built a 3D environment in which to view and interpret registered multi-modal data. Several different tools were developed and tested so that an analyst may be immersed within data and perform more advanced analyses. A few simplistic models were tested in order to demonstrate some basic capabilities, shown in Figure 2.5. In order to take full advantage of this environment it was necessary to develop a process to bring different signals together for analysis and obtain descriptive information of greater complexity than just the physical characteristics. For this thesis it is assumed that an AANEE environment with 3D registered multi-source data, models of the plant, and process simulation (for estimations/simulations of variables) is fully functional. This is used to understand how new analysis paradigms might be facilitated in such an environment. In many cases both data and functionality may need to be simulated since the goal is to explore new approaches long before full capabilities exist to help determine if this is a productive approach.

2.2 Site Identification

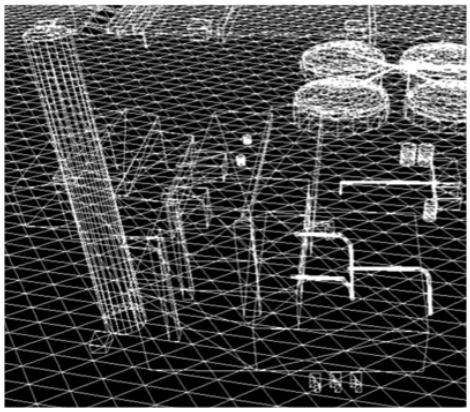
For this study it is assumed that the physical location of a site has already been determined before one would even begin this process. For the purposes of this



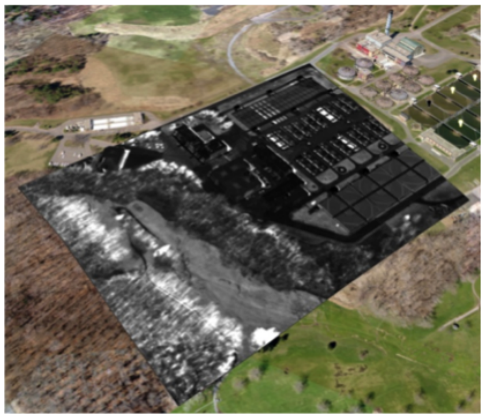
(a) 3D viewing enabled with red and blue shifts.



(b) Lightbulb icons indicative of a pass/fail clarity test of the settling tanks.



(c) A wireframe mesh of the site.



(d) An IR image projected on to the terrain.

Figure 2.5: *A few examples of the capabilities of the AANEE analysis environment.*

project the term 'site identification' is referring to identifying all of the pieces of the site, both what the different objects are and what they do. More precisely: the interrelationships amongst the buildings and infrastructure that are isolated or demarcated by a security perimeter. This is done in two steps. First is *infrastructure analysis*, where the functions or processes or individual structures are identified. The other step is *process identification*, where the main function of the facility is determined by putting the pieces together.

2.2.1 Infrastructure Analysis

When one encounters a site for the first time it is often easiest to investigate the layout of the buildings and roads on the site first, as these are typically available for viewing from airborne imagery. Deciphering the main function of the plant, however, may prove difficult without knowing all of the input and output materials.

Buildings, Roads, and Machinery

What is often the most accessible information through traditional remote sensing methods is the site infrastructure. Buildings, roads, pools, and electrical transformer yards are large objects that are not easily hidden and change very little over short periods of time. Many clues about a site are evident from its infrastructure. Different types of facilities will need different types of cooling towers, in different numbers, and possibly in multiple locations. Others might need pools of water or other chemicals for various processes. One can get an idea as to the number of workers on the site and the number of different shifts by counting the number of cars in the parking lot throughout the day. The two sites shown in Figures 2.6 and 2.7 provide examples of how the physical objects can provide many clues about facility operations.

Figure 2.6: *This is a rather large plant which has countless vehicles all around. There is also a closed road course on the left side which seems to indicate this is a vehicle manufacturing facility. Image courtesy of Google Earth.*

Figure 2.7: *A relatively small facility with a giant cooling pond. The large electrical transformers give away that this is a power plant. Image courtesy of Google Earth.*

Buildings are functional objects; they are rarely put up without knowing what is intended to go inside. Using this information, one can assume that each building often represents one or more different phases of the process taking place on a site, like the example shown in Figure 2.8.

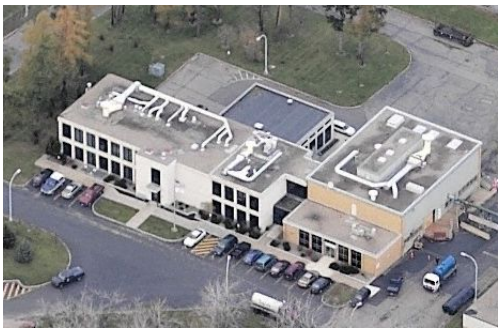


Figure 2.8: *A building with several windows, cars parked out front, and various pipes and vents on the roof. This is an administration building with some small scale gas or chemical testing taking place within. Image courtesy of Bing Maps.*

Multiple buildings of similar dimensions often implies that the buildings house the same equipment, such as in Figure 2.9. An analyst can use this information to figure out how many phases there are to the process running at the site and compare the buildings to those on known sites. This helps the analyst draw an accurate conclusion as to the purpose of the facility on a target site (Allen, 2008).

Figure 2.9: *Two identical looking circular tanks with several pipes, performing the same process in parallel or similar sequential processes. Image courtesy of Bing Maps.*

Some buildings are made to house large machines. These machines are often run

with very large motors. Motors and engines give off radio waves while they run, and they are often fairly unique to the type of motor. While these signals are unlikely to be picked up from airplanes, one could envision a passive RF detector placed on the ground nearby, potentially reading the signals given off by the motors. These signals should help clue an observer as to how many motors are running at a given time within the site of interest. Motors generate heat while in operation. Should the motor be large enough, or should there be enough of them in operation, the heat will be significant enough to be able to detect the activity from outside the building.

The facility in Figure 2.10 is not as obvious to classify as the sites shown in Figures 2.6 and 2.7. There is either one giant building or several smaller buildings that are all connected; the only thing that is clear is that there are definite points of separation along the roof showing they are different ages, materials, or both. To get a sense of scale, compare the size of the buildings with the size of the cars in the surrounding parking lots. Every segment is quite large, indicating some very large machinery must be inside, but without more information it is difficult to determine what processes are taking place within. With more historical data it might be possible to identify when each segment was constructed, and possibly track the changes in terrain to see how much earth was moved with each addition. More insight can be obtained by examining the input and output materials in order to help identify such facilities.

Figure 2.10: *A facility with large buildings and ample vehicles, indicating a manufacturing facility of some kind, but of what is not easily determined without more information. Image courtesy of Google Earth.*

Input and Output Sources

A site of interest is similar to a mathematical function in that it performs some operation on its inputs to produce an output. Input sources are often numerous, and each one may need to be accounted for. One can watch the outside of a site and count deliveries as they are being made and get key information such as quantity and frequency. Most industrial facilities will need a source of water, electricity, fuel, and waste outlets. Most modern sites also require cell phone service and high-speed internet access (Commercial Real Estate, 2011). Unfortunately, most of these are typically done underground making them difficult to detect, though not always impossible as demonstrated by Figure 2.11. The physical demands a facility has on its environment are often key clues as to the exact process taking place within. It is also possible to track shifts of the employees working at a site, allowing one to become familiar with how many people are working during each shift of the day.

Figure 2.11: *A thermal infrared image showing hot underground pipes (bright white lines). Image courtesy of Petrie, 2001.*

For most sites the dominant output is the primary function of the facility. But there are multiple types of of output that can, like input, take several forms. Some products likely have to be physically taken off site, and it is not always easy to discern the nature of these products. Some might be waste that is being taken to a dump, it might be a byproduct of the site's internal processes that is usable by some other site, or it could be a product the site is designed to produce. Any site could

produce any or all of these, and the method in which these products are taken off site will often vary. A site can also have output in the form of a gaseous or liquid plume, potentially released underground. Many types of facilities have cooling towers or smoke stacks and it is often of interest to know how often these towers are active and to what degree.

Once all of these have been collected it is often possible to determine the main function of a site as well as get an idea as to the size of the site. Size is referring not to the physical size of the site but the quantity of output and customer base being serviced. For example, one might be able to determine that a site is in fact a nuclear power facility and it is responsible for approximately 250,000 homes and businesses. In the circumstance that the site is not friendly it is unlikely that one would be able to get an intimate look at the input and output of a site. In such a situation one will have to rely on various remote sensing techniques. Underground gas and liquid plumes have to end somewhere, and are likely to have some effect on the environment, while liquids and gases on site can be monitored with spectral imagery. One could get more information on materials being brought to and from a facility with trucks by following them on the ground or with UAVs.

2.2.2 Process Identification

After thorough analysis of the buildings and other physical elements of the site it is important to determine their interrelationships and identify the key process(es) taking place therein. Even if the main function is known, one will want to analyze every structure and build a flow diagram of the various materials as they travel through the site, and note the different processes taking place at each location. It is important to make sure that nothing is missing, and that there are not any extra buildings or pieces of infrastructure. In order to truly know a site one must be familiar with the role of every component (Schneider, 2011). This is necessary because it helps one to determine the alternative processes that could be taking place. One does not analyze a toy factory because it makes toys, but more to know what chemicals are being used to make those toys, in what manner, and are they

being used appropriately.

Say an analyst learns that the facility in Figure 2.10 regularly has rolls of white paper dropped off and magazines picked up. Logically, one would infer that the site is actually a printing press (it is actually the Quad Graphics plant in Saratoga Springs, NY), and identifying the facility with this process provides a lot more information than Paper → Printing → Magazines. Ink and silicone are applied to the paper, strong cleaning chemicals are required to clean the ink rollers in between jobs, and various adhesives are needed for the binding process. It is up to the analyst to use this information to figure out what types of alternative processes could be done with these various chemicals and determine the signals that would manifest in those situations.

2.3 Observable Signals

2.3.1 Data Collection Methods

Once a site has been named, its possible processes determined, and its different operational modes identified it may now be possible to determine the signals that will help an analyst differentiate said modes. The observable signals are linked to the different modalities we are identifying. Subtle differences that help identify varying amounts of production are clues that might aid one in determining the difference between a chemical production plant that is doing as it claims, or secretly making chemical weapons. There are certain key pieces of data that can and need to be observed in order to tie everything together into a single coherent package. An in depth discussion on specific observables for the Van Lare site used in this study is included in Chapter 4. What follows here is a discussion on the different methods by which data can be collected and some examples of the multiple forms of data that each method can collect. This thesis seeks to describe a generic process with an extensible architecture so that different forms of data can be incorporated as they become available.

Overhead Imagery

All overhead imagery is, of course, a function of the ground sample distance, or pixel size, of an image. This project assumes reasonable resolution is available in all formats in a fully registered, interactive AANEE environment. From basic overhead RGB imagery it will be possible to detect a large quantity of the desired information. When first given a new site of interest the first thing almost any analyst does is look at the panchromatic or RGB imagery to get a feel for the site design and infrastructure. Through this it is possible to get the first estimation of site operations and to locate key points of interest such as entry and exit points as well as delivery stations and administrative buildings. Specific to the Van Lare plant it is possible to determine the number of aeration tanks and settling tanks in use, as well as construction areas where something is either being repaired or upgraded.

Through hyperspectral image data it is possible to learn much more about the site of interest. Different materials have different spectral properties and it is often possible to determine the make-up of the roof and walls of a building. The physical make-up of a building, as well as its shape as mentioned before, provides information as to the purpose of the building. The visible and short-wave infrared regions of the spectrum are dominated by reflected radiance, and would need to be collected for material identification only every time a new building or other construction object is added to the site (Schott, 2007). Hyperspectral data also allows an analyst to find gas plumes and determine constituents in ponds or pools (O'Donnell, 2005). The Van Lare site has several open settling tanks and aeration tanks, as well as a chemical mixing tank. The final settling tanks are supposed to be mostly clear but this is not always the case. At times equipment can be broken or chemical and biological processes may not be taking place in optimal conditions. This can lead to the quality of the effluent water not being as high as normal standards (Bartlett, 2007; Lukas, 2007).

Knowing the relative temperature of various buildings, transformers, and other infrastructure is essential for noticing the subtle differences between operational modes. LWIR imagery was utilized in multiple forms for this purpose.

Oblique imagery (like the images available on Bing maps) is useful in bringing the analyst information that is available on the sides of the building but cannot be detected from the roof. Such incidents occur when a large machine is running on a lower level and the thermal signature does not reach the upper floors (Schott, 2007; Pictometry, 2007). Historic imagery could also prove to be quite useful. Old images can show when a building is being constructed or a new piece of equipment is being installed. One might see pipes or conduits from building to building or tank to building. One might see the depth of an excavation before a tank goes in, and then compare that to the depth of the tank once installed but still not in use or empty for repairs.

Remote Ground Detection

Through ground based surveillance around the perimeter of a plant additional types of information can be recorded. Such reconnaissance can yield data in multiple formats. One could obtain ground-based obliques of several buildings and transformers on the site, in the visible or in infrared.

Physically leaving a person on the ground outside of a plant to record information is the best way to track the vehicular traffic at a site, however, in a real situation where the site is potentially in hostile territory, such surveillance is not possible. In order to keep with the idea that this is a proof of concept project and not made specifically for the Van Lare site it is necessary to consider all scenarios and base the process off of the most difficult. It is possible, therefore, to imagine a situation where a sensor is monitoring the entrances into a site. Most sites with any security have very few points of entrance and exit. A wide angle lens could be used to provide additional coverage, as shown in Figure 2.12.

Getting a grasp of the vehicular traffic is important for getting a better understanding of site operations. Through this, it will be possible to estimate the employee quantity and scheduled shifts. More importantly is tracking the industrial deliveries, pick-ups and drop-offs, that make their way onto the site. An analyst will want to have a good estimate of the frequency of each type, as well as the quantity and

Figure 2.12: *Fish-eye view of the entrance to Van Lare.*

nature of the visit. Specifically to Van Lare, it is known that wastewater treatment plants use chemicals to treat the water to kill bacteria and keep the PH balanced. There are also trucks that come to pick up treated sludge, septic companies that empty their trucks, and restaurants that unload their "scum" - greases and oils that were used in cooking (Lukas, 2007).

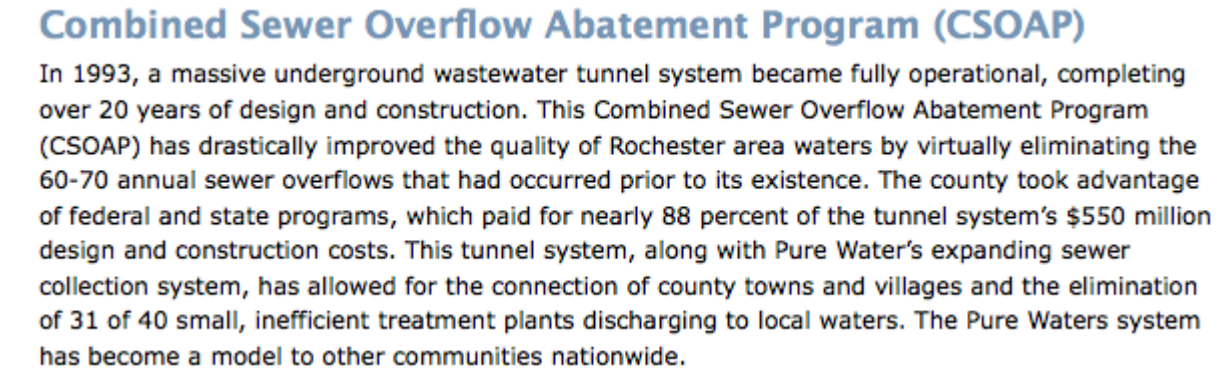
Most large facilities utilize hand radios for employees to maintain communication. During normal activities there should be an average level of communication that is necessary to keep the facility running properly. Should a random incident occur, radio traffic is likely to increase for a brief period of time in order to notify all plant employees of the situation and what needs to be done to bring it back to normal. Passive radio receivers are cheap and can be set up outside a facility to monitor such traffic.

Mass Media Intelligence

It is always important to take advantage of people that are doing work for you. Reporters and journalists have the general responsibility of keeping their eye on the world and notifying the public about interesting events. Through Internet, newspapers, public records, and television it will be possible to get valuable information on

or in the area surrounding a site of interest.

Underground improvements that take place within the city of Rochester are not things we are going to be able to see through typical surveillance methods. These activities are taking place a great distance away from our site of interest, so it is unlikely we would even have a sensor there to detect these changes. These are important to note, however, as improvements to the storm drainage system or sewer system have a direct impact on Van Lare operations and were likely initiated by plant employees or county employees. In the article shown in Figure 2.13 we can see that there was a large city wide project taking place that eliminated several small wastewater treatment facilities in the 1990s. This caused significantly more wastewater to be directed to the Van Lare treatment plant. The facility had to put in the large settling tanks, shown later in Figure 3.9(b), to prepare for the increase in volume.



Combined Sewer Overflow Abatement Program (CSOAP)

In 1993, a massive underground wastewater tunnel system became fully operational, completing over 20 years of design and construction. This Combined Sewer Overflow Abatement Program (CSOAP) has drastically improved the quality of Rochester area waters by virtually eliminating the 60-70 annual sewer overflows that had occurred prior to its existence. The county took advantage of federal and state programs, which paid for nearly 88 percent of the tunnel system's $550 million design and construction costs. This tunnel system, along with Pure Water's expanding sewer collection system, has allowed for the connection of county towns and villages and the elimination of 31 of 40 small, inefficient treatment plants discharging to local waters. The Pure Waters system has become a model to other communities nationwide.

Figure 2.13: *An example of mass media intelligence. The article is available at http://www.monroecounty.gov/des-index.php.*

Several other events are likely to take place on or around a facility, some of which may be relevant to the site, but many that are not. For example, anytime there is a smell complaint made by residents near Van Lare a full report has to be made and documented by county officials (Lukas, 2007). When a wastewater treatment plant is running properly there should not be any odor (Bartlett, 2007). Obviously a complaint made by local residents about the odor implies that something is not

going according to the standard operating procedure. One would have to look at the time delay that occurs from the actual incident to the time when the actual report is being made in order to determine the cause. There are several possible explanations for odor to become a problem to the residents. A malfunction in the air purifiers that are used to remove the odor from the air or too much volume and an inability to provide everything with the necessary treatments, for example. A specific incident that was described by a plant employee was determined to be caused by high winds.

Other documented events that may have an effect on a plant are criminal activities in the area. Muggings, murders, robberies, and vehicular accidents are news worthy events that could potentially influence the activities or operations at various sites of interest, so these occurrences should be noted, should they happen, for completeness of the project. A vehicular accident involving an industrial vehicle en route to or from a plant is likely the most relevant event to potentially take place, but it would have to be a catastrophic event in order for it to have any real impact on the plant.

Some industrial facilities are dependent on the weather, and this is likely the most important and influential piece of information from mass media intelligence for Van Lare. While Van Lare is in a relatively stable weather area and only affected by rain and snow, other facilities might have to compete with violent thunderstorms, tornados, hurricanes, or earthquakes. These events can lead to shut downs, spills, or even explosions. An analyst may be monitoring a plant simply to see if it is capable of standing up to such a catastrophic event.

On Site Measurements

While on site measurements do not qualify as remote sensing they do provide valuable information used in process identification. These can be obtained in the event that UN investigators or some other regulatory organization were ever to investigate a site of interest or by taking measurements from a similar site. There are finite methods available to obtain an end product given the ingredients. For example, to become more familiar with Van Lare it was beneficial to visit the wastewater treat-

ment plants in Geneseo, NY and Cazenovia, NY. While Van Lare is significantly larger than the other two, the inputs, processes, and outputs are the same. Van Lare does take a few extra steps but this is likely due to the inability of the processes to scale up to the level needed at this plant (Bartlett, 2007; Lukas, 2007).

Other measurements that might prove to be useful are measurements of products as they traverse through a site. It could prove useful to have concentration information of the different ingredients at each stage of production to get a good estimation of how much of each chemical is used and see if this correlates correctly with how often they receive a delivery. As with Van Lare, biological concentrations may also prove useful to know.

A collection of some of the different signal types and the data they can collect is available in Table 2.4. All of these could combine to give a detailed description of a facility, as well as determine all of its functions and operational modes. Fortunately there are several pieces that overlap, as it is highly unlikely that one would have access to all of these modalities at one time. For example, major environmental incidents such as a contaminant found in some nearby water source would likely be found with some spectral imager as well as reported in some form of media. Unfortunately the only way to verify everything taking place within a sealed building is to have an on-site measurement team. Such things are rarely allowed in most areas of the world, but analysis of a surrogate site could also yield promising results. Employee and facility records may be altered at a nefarious site, but if one can locate where or when the documentation has been altered it could help in determine exactly what a site is trying to hide (*e.g.*, a high level of operation during a time when records indicate limited activity).

2.3.2 Obtaining Real Data

One needs to determine exactly what it is one is looking for and identify the observable signals that can aide in obtaining this information. A person can then determine which sensor can be used to record the necessary signals. From there it becomes possible to calculate the numerical values the sensor will provide while observing

Table 2.1: *A collection of the possible overhead imagery data types and examples of what they could be used to detect at an industrial site.*

Overhead Imagery	
RGB	Perimeter identification and status Infrastructure identification Infrastructure on/off Number of vehicles
VNIR	Calm vs. turbid water
Thermal	Power usage Heat sources
Spectral	Gas plume detection and identification Material identification Constituent retrieval from liquids

Table 2.2: *A collection of the possible remote ground detection data types and examples of what they could be used to detect at an industrial site.*

Remote Ground Detection	
Oblique Thermal	Power usage Active pumps
Passive RF	Communication monitoring Number of motors in operation
RGB	Imports and exports Perimeter security
Spectral	Gas plume detection and identification Material identification Constituent retrieval from liquids

Table 2.3: *A collection of the possible mass media intelligence data types and examples of what they could be used to detect at an industrial site.*

Mass Media Intelligence	
TV, Radio, or Internet News	Environmental reports Increased crime Major incidents New infrastructure
Public Records	Budget items List of employees (government)
Weather	Rain, wind, and temperature measurements Major incidents

Table 2.4: *A collection of the possible on site measurements and examples of what they could be used to detect at an industrial site.*

On Site Measurements	
Infrastructure Inspection	Check for minor leaks Verify infrastructure relationships Collect samples
Facility Records	List of employees Accounting Pick-ups and deliveries Other recorded information
Physical Properties	Material quantities Processing rates

the signal and what it means. All of the items shown in Table 2.4 are quantifiable and will have a range, a mean value, and a variance. One can gather the signals from a collection of these items together as observation vectors. As vectors they can be computationally analyzed and used to aid analysts in drawing conclusions about a site of interest. Some things, such as the number of empty settling tanks, are easily quantified. Other things, like the ratio of the signal from the transformers to their surrounding area, are not as easily quantified without an understanding remote sensing systems. A brief discussion of the nature of these systems follows in this section.

Remote sensing, as it is discussed in *Remote Sensing: The Image Chain Approach* (Schott, 2007), deals mainly with the collection of imagery through the use of airborne instruments, which is what is typically being referred to when traditional remote sensing approaches are mentioned. There are four key components to sensor reaching radiance. Solar reflected radiance (L_s) is that which comes from the sun directly to a facet, and is reflected to the sensor. Downwelled radiance (L_d) is that which comes from the sky and reflects off a facet to then go to the sensor. Solar and downwelled are often combined and referred to as reflected radiance (L_r). Emitted radiance (L_e), which is more dominant in the long wave infrared (LWIR) region of the spectrum, is that which is naturally given off by a facet or object. And lastly, upwelled radiance (L_u) is that which comes from the sky and goes to the sensor. In short, we have

$$L_{Total} = L_r + L_e + L_u \tag{2.1}$$

and a more in depth explanation can be found in Chapter 4 of Schott, 2007.

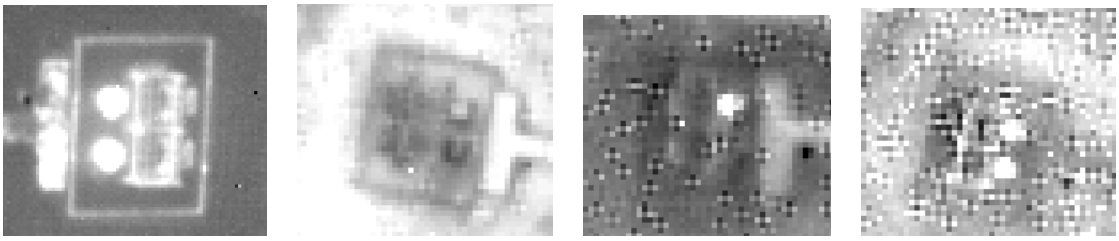
For example, a settling tank can either be full or empty. If the settling tank is full of water then the light hitting that water will be scattered within and, if deep enough, very little light will be reflected from the bottom. This will cause a sensor looking at that tank to record a very low signal thus making the tank appear dark. On the other hand, if the tank is empty and the gray concrete on the bottom is visible then this concrete will reflect a large amount of the incident light, thus causing the sensor to record a high signal and making the tank appear comparatively

brighter. Obviously most things are more complicated than that so a more complex case is examined next..

The aforementioned LWIR band is also known as the thermal band and can be used to detect processes taking place inside of walls and underground. All objects emit (L_e) photons in this region that is proportional to their temperature. Extremely hot items will show up bright (like two of the transformers in Figure 2.14) while extremely cold items will give off little to no signal, thus showing up as dark. Upon visual inspection of LWIR imagery of the transformers it is fairly easy to notice that Jan 18, 2007 and May 3, 2008 have two bright (hot) spots while August 1, 2007 only has one. In the July 24, 2007 image one could argue that there is a bright-ish spot over one of the transformers but it is not entirely clear until one looks at the actual digital count values in the image.

Figure 2.14: *LWIR image of the Irondequoit pump house at the Van Lare facility and a transformer yard (outlined in red) that is directly related to the power usage of the building.*

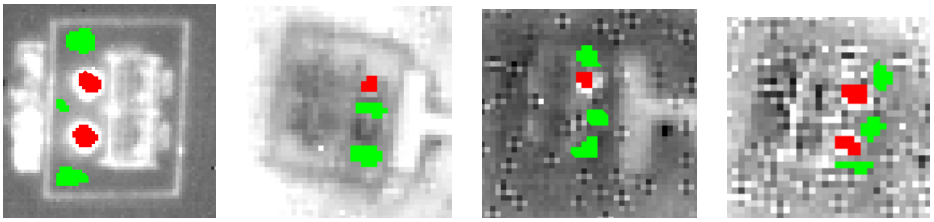
The average value of the bright spots can be calculated and compared to the average pixel values of the regions around the transformers (shown in Figure 2.16). As can be seen from the values shown in columns two and three in Table 2.5 the values change dramatically, likely due in large part to the air temperature. In order to get a more consistent measurement one can look at the ratio of the values, as shown in the fourth column of the same table, or the difference in digital counts,



(a) Jan 18, 2007 (b) July 24, 2007 (c) August 1, 2007 (d) May 3, 2008

Figure 2.15: *LWIR images of the transformer yard.*

shown in the fifth column, or the percent increase, as shown in the final column.



(a) Jan 18, 2007 (b) July 24, 2007 (c) August 1, 2007 (d) May 3, 2008

Figure 2.16: *LWIR ROIs of the transformer yard.*

Table 2.5: *Ratio of LWIR signals, the difference in digital counts, and the percent higher of the transformers to the area around them.*

Date	Avg Red	Avg Green	Ratio	Difference	Percent
1/18/2007	4017.26	3410.47	1.178	606.79	17.8%
7/24/2007	6455.71	6294.3	1.026	161.41	2.6%
8/1/2007	7291.14	7045.15	1.035	245.99	3.5%
5/3/2008	6128.90	5880.38	1.042	248.52	4.2%

What has been demonstrated here is while many types of signals are easily analyzed and compared, sometimes one or more simple tricks will have to be implemented first to provide the analyst with comparable data. Also it should be becoming more clear exactly how complex the problem is that this thesis attempts. There are dozens of observable signals, each with its own range of values, all of

which are to be collected over time to determine what a site of interest is doing. It is important to note that, as a result of this project, one will not be able to say with absolute certainty what is happening at a particular site. It only hopes to bring to light the process by which this analysis should be done so that the intelligence community can improve its probability of success.

2.3.3 Summary

Using the methodology described in the last few sections one can put together a long list of information pertaining to a site of interest as well as make intelligent guesses as to the activity within. The major goal of this project is to provide a method for analyzing many types of signals simultaneously so that one can then take that information and use that to determine the current state of operation at the site. To do so requires information collection. Information collection not only means ordinary measurements and signal collection, but also gathering information about the relationships each of the different signals have with one another as well as their role in the process an analyst is trying to detect.

2.4 Data Interpretation

A single piece of data is typically indisputable. A light is on or it is not, car is moving or it is not, a pipe is leaking or it is not. Once different types of data start being put together for complex analysis the clear perspective becomes blurred. The data is now open to interpretation, and the manner in which it is interpreted will vary by the method that was used to analyze it. Like two people reading the same poem, it is unlikely that the conclusions drawn will be complete opposites of one other, but small variations are almost guaranteed to manifest themselves. Regardless of the method used to analyze data, some things just cannot change. Given three observables with different types of signals, the mean and variance of each individual signal will not change, nor will the covariance or correlation. The things that can change are the manner in which these are used. What follows in this section is a brief

discussion of three simple methods that can be used to analyze high dimensional data, namely template matching, geometric analysis, and statistics.

2.4.1 Types of Metrics

This section describes three methods that are used to compare and analyze high dimensional data, as well as demonstrates the manner in which they apply to this project. The effectiveness of each is demonstrated in different situations in Chapter 4. First one must realize that there are some completely different cases in which this analysis can be done. Four such possibilities are listed below with examples (illustrated in Figure 2.17) (Sipser, 1997).

1. Disjoint states that make up the only possible states. This is possible when one has a very specific goal in mind. For example, is this site making a biological weapon. Either they are or they are not.

2. Disjoint states in a space with an unknown number of other states. This happens when one is doing a more general search, like what kind of disaster is taking place here. There might be a couple that are clearly known and searched for, but there are plenty of other possible things that could be going wrong.

3. States with some overlap in a space with an unknown number of states. Possible if an analyst is trying to determine the operational mode of a site.

4. States where it is possible to have one be a complete subset of another in a space with an unknown number of states. This happens when it is possible for one or more variables to change the manner in which the operational modes of a plant are being done.

Disjoint states are the simple states of a settling tank. The settling tank either has water in it or it is empty. There is no overlap and there are no other possible states. Using the entire plant, this binary case becomes ternary: high, medium, and low flow operational modes.

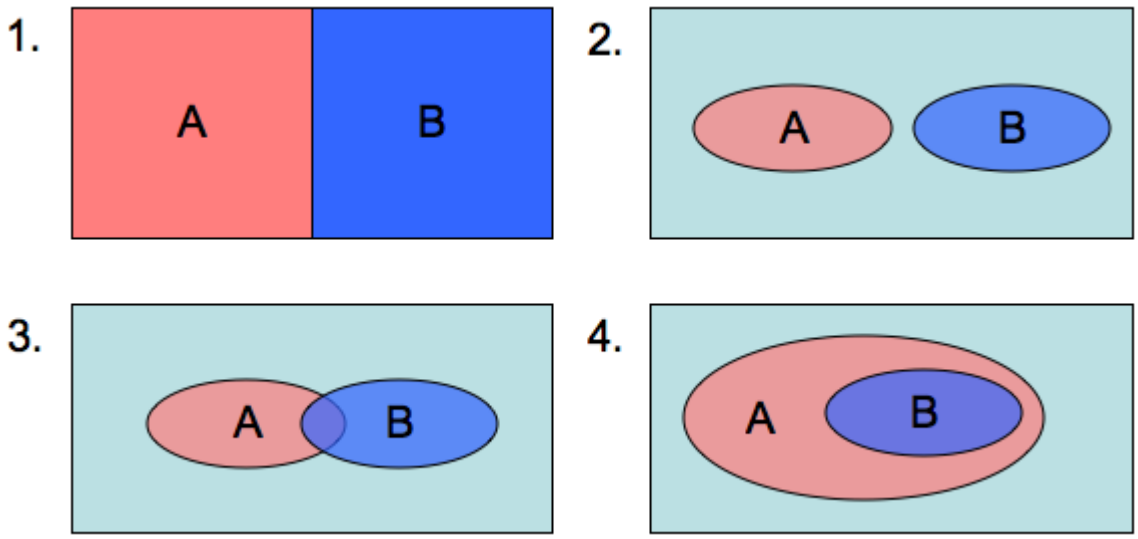


Figure 2.17: *Venn diagrams of four cases using only two states.*

Further examination of the settling tank shows that there are other possible states. Being full does not necessarily mean it is on. Being empty could mean it is broken or that it just does not have wastewater in it at the moment. Again relating that to the whole plant we know that if it is running it is in one of the three operational modes. However there are several other possibilities, such as a potential maintenance mode, a failure mode, and even a shutdown mode. Just because these events are not being searched for does not mean they do not exist.

Some states could have some overlap, especially since the differences in the operational modes could be arbitrary lines in flow rates. For example, if low flow mode is less than 90 million gallons per hour and above that is medium flow mode, flow rates very close to that mark will have observables that are very similar, if not identical. This implies that there might be a bit of overlap in the plant states.

Case 4 is unique in that it does not apply directly to the main operational modes, as in low flow mode is not a subset of high flow mode nor is high a subset of low. There are, however, other interesting things happening at Van Lare. There are two possible sources of wastewater input, different types of phase one settling tanks, and

various other scenarios like the potential failures mentioned above. High flow mode with most of the wastewater coming from Irondequoit and some broken east side settling tanks will have some different signals than high flow mode with wastewater coming primarily from the city with all the settling tanks working properly. These two occurrences are subsets of high flow mode, but they could also occur during medium and low flow modes. Another way to look at them is as flavors of the operational mode, just like ice cream cones can come in one, two, or three scoops and each size may be vanilla, chocolate, or strawberry (Peebles, 2001)

Template Matching

Template matching is the simplest method demonstrated in this thesis. Given that there are two states, A and B, and n observables, then one can create two $n \times 1$ vectors $\vec{A}$ and $\vec{B}$ such that $\vec{A} = \vec{B} \iff A = B$ (*vector A equals vector B if and only if state A is equal to state B*). Now one can introduce a test vector $\vec{C}$ that contains all of the n observations collected at a single point in time. The definition of "single point in time" will depend upon the site of interest and could easily vary from a microsecond on up to weeks or months. For this project a single point in time is often referring to 2-4 hours. Two methods are used to determine which state $\vec{C}$ is most representative and each method needs to handle each of the four cases shown in Figure 2.17 (Sispser, 1997).

There are multiple methods that will be used to weight the different variables in order to calculate the probabilities. In many situations all variables will be weighted the same, as if they all provide the same amount of information in determining the result. It may also be of interest to weight variables based on how correlated they are with the mode being predicted. In some cases it may become necessary to weight variables differently based on whether or not a signal is detected. As an easy example, if an explosion is present at a facility, something is going wrong, where an explosion not being present does not clearly indicate that everything is perfectly fine. Lastly, the variables will be weighted based on how reliable they are in mode prediction.

One of the ways to weight variables based on reliability is through the application of Dempster-Shafer theory. A reliability for each variable can be determined by calculating how often a variable accurately predicts a mode. When predicting the mode of a facility, all of the variables are then combined together. Often times some variables will suggest one mode, while others are indicative of another mode. Dempster-Shafter theory calculates the probability of each mode by comparing the likelihood several variables being reliable and a few being unreliable to both cases of several variables being unreliable with some being reliable and all variables being unreliable (Sentz, 2002). This is demonstrated later in this section in Table 2.6.

$\vec{C}$ in Case 1 There are two ways to check to see of which state $\vec{C}$ is a member. In a singular matching case each observable c_i in $\vec{C}$ will be compared to its associated observable a_i in $\vec{A}$. If c_i is either equal to a_i or at least within some predetermined threshold then state A will get a point. This process is then repeated for $\vec{B}$. $\vec{C}$ then is a member of the state with the highest score as shown in Figure 2.18. This is similar to a doctor trying to find out what is wrong with a patient by comparing the patients symptoms to known symptoms of an affliction one at a time (OpenClinical.com).

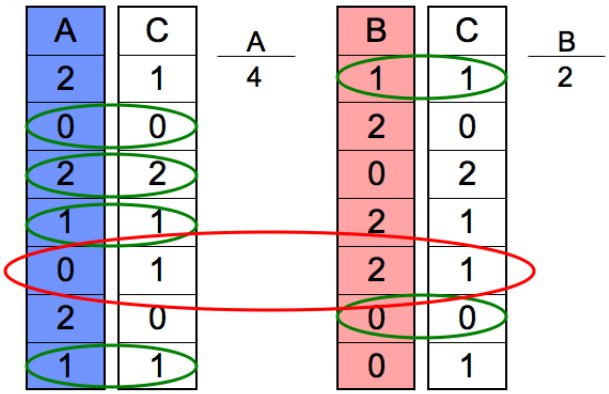


Figure 2.18: *Observation $\vec{C}$ is being compared to templates $\vec{A}$ and $\vec{B}$. In this case each value of $\vec{C}$ must be within thresholded proximity that is set at 0.5 to be considered a match to its associated value. The row circled in red shows that observation did not match either template at that point.*

An alternative method is to compare the elements of $\vec{C}$ with the elements of $\vec{A}$ and $\vec{B}$ simultaneously. This is more like the method of differential diagnosis performed by doctors whenever they are with a patient with an unknown ailment (WebMD, 2011). In this method one simply determines which value (a_i or b_i) c_i is closest to and gives a point to the associated state. Again, $\vec{C}$ is said to be in the state with the highest score. In Figure 2.19, $\vec{C}$ would be said to be a member of A.

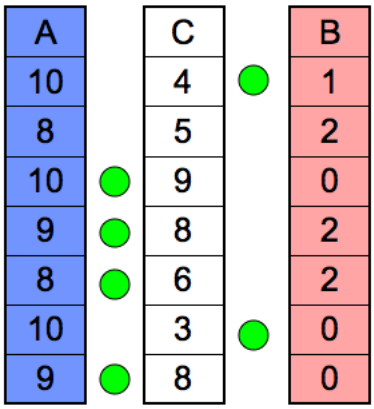


Figure 2.19: *Comparing $\vec{C}$ to templates A and B. In this case a match is determined based on the relative closeness of each element of $\vec{C}$ is to the corresponding template elements.*

$\vec{C}$ in Case 2 In case two $\vec{C}$ can only be tested to be in states A or B similarly to the singular matching method in case one. There are, however, two ways in which to interpret the results. For example, one might require a state to achieve a minimum score in order to be considered a member of that state. If one has $n = 7$, perhaps one would say that $\vec{C}$ is in A if A receives a score of at least 4. In a more strict scenario one could require a score of at least 6, and in that situation $\vec{C}$ from Figure 2.18 would not be a member of either state. The threshold will be dependent upon the site, the observables, and how damaging it can be to have false alarms. If $\vec{C}$ passes the threshold to be a member of both states even though it is known that they are disjoint then the threshold must be increased.

The other way in which to analyze the results in this case is to assign a probability

that $\vec{C}$ belongs to either state based on the score. Again, assuming that $n = 7$, perhaps A gets a score of 4 and B gets a score of 2 as shown in Figure 2.20. In such a situation one could say that $\vec{C}$ is 57% likely to be a member of state A, 29% likely to be a member of state B, and 14% likely to be a member of neither state.

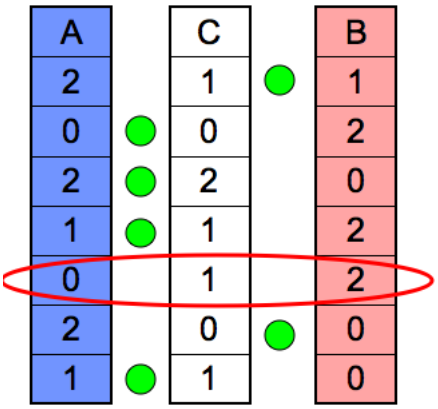


Figure 2.20: *Comparing $\vec{C}$ to templates A and B and determining the probability of each state based on matches. With one row producing no matches there is also a chance that $\vec{C}$ is in neither state.*

$\vec{C}$ in Case 3 If A and B have some intersection one will want to use the second method described in case two, and illustrated in Figure 2.20. What is unique about this case is that $\vec{C}$ could be a member of both A and B.

$\vec{C}$ in Case 4 As mentioned before the subsets are more like flavors of each mode. So while each of the three operational modes is either disjoint from the other sets or has some small amount of overlap, the different flavors can be applied to each mode. Thus each state is then declared to be its own space that can again be subdivided into several new states (flavors) as shown in Figure 2.21.

Figure 2.21 shows two groups of states. A, B, and C are three major states. A is then expanded to show that it is its own space with sub-states 1, 2, and 3 within it. These sub-states can be parts of each of the primary states. Using the Van Lare facility as a reference there are three easily identifiable modes: low, medium, and

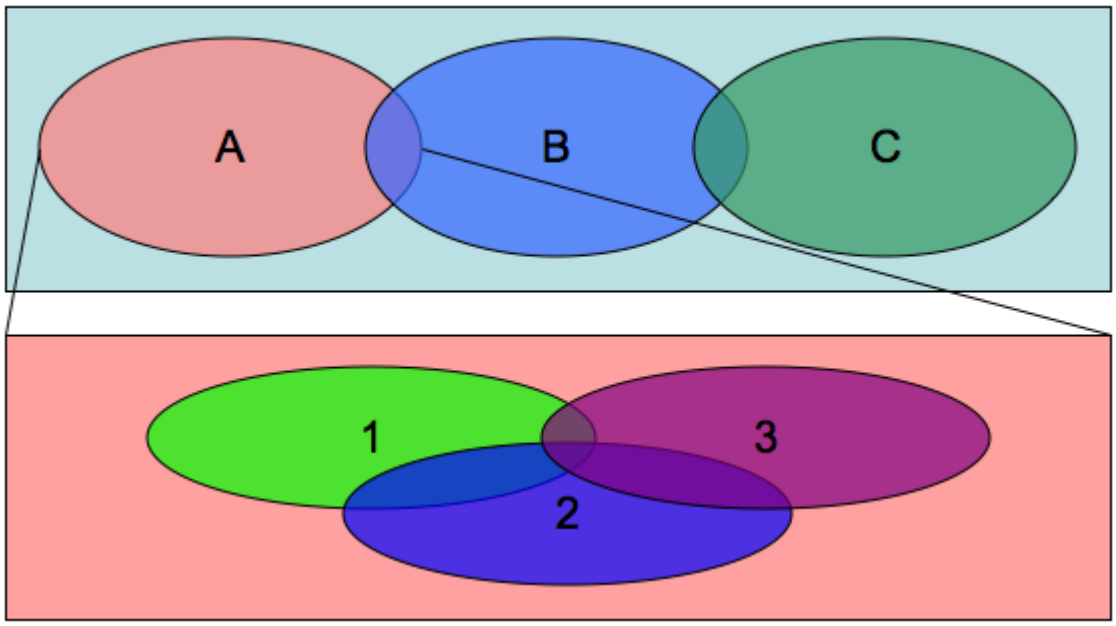


Figure 2.21: *A representation of each state as its own space with sub-states.*

high flow modes. There are varying signals from the observables that differentiate these modes, but there are also observables that are completely independent of the operating mode and simply change the flavor of the mode. These flavors are often picked up through more subtle clues, such as the thermal signal off of the transformer yard outside of the pump house. If this signal is very high while the plant is running in medium flow mode then one can infer that there is more wastewater being treated at the moment from near by Irondequoit, NY rather than from the city of Rochester. It is not assumed that this research will make such things perfectly clear, it will hopefully bring to light how capable one would be at making such observations.

Dempster-Shafer An alternative method to calculating the probabilities is through the use of Dempster-Shafer theory. Using the data from Figure 2.19 there are seven variables. In order to use Dempster-Shafer one must assign a reliability to all variables. For the sake of this initial example assume they are all 90% reliable. In this case there are four variables saying that $\vec{C}$ is a member or class A, two variables

saying that $\vec{C}$ is a member of class B, and one that is abstaining. Obviously, in this situation it is not possible for all of them to be reliable since there are conflicting results, so the possibilities are that four of the variables are reliable and two are not $(0.9^4 * 0.1^2)$, two of the variables are reliable and four are not $(0.9^2 * 0.1^4)$, or all six of them are unreliable (0.1^6). One can then determine the probability of each of these cases by summing up those results and dividing each product by that sum. This is illustrated much more clearly in Table 2.6 (Sentz, 2002).

Table 2.6: *An example of the application of Dempster-Shafer theory. Here it is assumed that all of the variables have the same reliability, and the probability of each case is shown as the reliability changes.*

Reliability	4 Rel/2 Unrel	2 Rel/4 Unrel	6 Unrel	P(A)	P(B)	P(Neither)
.9	$0.9^4 * 0.1^2$	$0.9^2 * 0.1^4$	0.1^6	98.8%	1.2%	0%
.8	$0.8^4 * 0.2^2$	$0.8^2 * 0.2^4$	0.2^6	93.8%	5.7%	0.4%
.7	$0.7^4 * 0.3^2$	$0.7^2 * 0.3^4$	0.3^6	82.1%	15.1%	2.8%
.6	$0.6^4 * 0.4^2$	$0.6^2 * 0.4^4$	0.4^6	60.9%	27.1%	12.0%
.5	$0.5^4 * 0.5^2$	$0.5^2 * 0.5^4$	0.5^6	33.3%	33.3%	33.3%

Geometrical Analysis

A geometrical approach to this problem involves plotting all of the data in a high dimensional space. In doing so, different clusters should manifest within this space. Each of the clusters should be representative of the different operational states of the plant. The vector C in this approach is a point in n space. In all of the following cases there would need to be enough data available to perform the testing. If there are not enough data collections to create a full test space then the data will need to be supplemented with simulations. Doing so will require fairly accurate probability distribution functions for all of the n variables so that the cluster centers of each mode are properly represented. Inaccuracies in these functions can lead to misclassifications or inaccurate confidence levels.

K-means is an unsupervised clustering method that uses an iterative approach to organize the data in a user supplied number (k) of classes. It has been described

as trying to minimize the function shown in Equation 2.2.

$$F(W, Z) = \sum_{l=1}^{k} \sum_{j=1}^{n} \sum_{i=1}^{m} w_{lj}(z_{li} - x_{ji})^2 \tag{2.2}$$

Here k is the number of clusters, n is the number of points, m is the number of dimensions, Z is the set of cluster centers, z is each individual cluster center, W is the set of probabilities that a point belongs to each cluster, and w is each probability value. A point belongs to a cluster when its distance to the cluster center is smaller than it is to the other cluster centers so W is all just zeros and ones. In the event of a point being exactly the same distance from two clusters it will be assigned to the cluster it was measured from first (Schott, 2007).

Once the cluster centers have been determined it is time to classify the test point C. This is done using the Mahalanobis distance of the test point C to the cluster centers (Z) which is calculated as shown in Equation 2.3. S is the covariance matrix of the points that belong to each cluster.

$$D(C) = \sqrt{(C - Z)^T S^{-1} (C - Z)} \tag{2.3}$$

When dealing with different types of data it is necessary to normalize the quantities to perform computational analyses. For the following examples all of the data have been normalized over the interval from 1-7 for ease of display.

After forming the array of distances it is time to begin forming the clusters. It is best to start with the two points that are closest together and combine them into a single cluster, so this is done with points 3 and 4 from the example. As shown in Figure 2.23 the distances to the new cluster from all other points are recalculated by taking the minimum distance of the two connected points to all points. This method would continue and the end result would be dependent upon which of the four cases one is assuming to be true (Kolodnikova, 2003).

	x	y
1	2	5
2	1	3
3	6	1
4	6	2
5	2	2
6	7	3
7	4	1

(a) 7 Points

	1	2	3	4	5	6	7
1	0	2.24	5.66	5	3	5.39	4.47
2	2.24	0	5.39	5.10	1.41	6	3.61
3	5.66	5.39	0	1	4.12	2.24	2
4	5	5.10	1	0	4	1.41	2.24
5	3	1.41	4.12	4	0	5.10	2.24
6	5.39	6	2.24	1.41	5.10	0	3.61
7	4.47	3.61	2	2.24	2.24	3.61	0

(b) Distance Array

Figure 2.22: *The points (left) and the distance each point is to all other points (right).*

	1	2	3	5	6	7
1	0	2.24	5	3	5.39	4.47
2	2.24	0	5.10	1.41	6	3.61
3-4	5	5.10	0	4	1.41	2
5	3	1.41	4	0	5.10	2.24
6	5.39	6	1.41	5.10	0	3.61
7	4.47	3.61	2	2.24	3.61	0

Figure 2.23: *After points 3 and 4 are made a cluster, the distances are recalculated.*

Case 1 In this case the number of clusters will be equal to the number of states for which one is testing. Using a two state scenario and the example started in Figures 2.22 and 2.23, the clusters would be broken down to the form shown in Figure 2.24. The center of the cluster is approximated by taking the average of the coordinates. One would then calculate the distance of C to the two cluster centers using Equation 2.3 If a test point is given such that $C = (4, 4)$ then C is 2.42 away from cluster 125 and 2.85 from cluster 3467, therefore it will be classified as a member of cluster 125 (Borgatti, 1994).

If one is allowed to alternatively provide a confidence level of each state then this

	Cx	Cy
125	1.67	3.33
3467	5.75	1.75

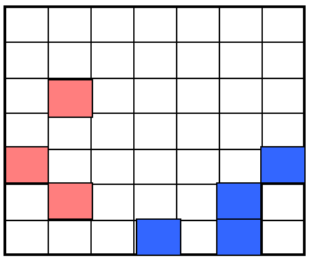


Figure 2.24: *Cluster results for case 1.*

can be done in a linear fashion by summing the distances C is away from each cluster and calculating $(MaxDistance - distance)/MaxDistance$. In this case that would mean that one is 54.08 percent certain that C is in cluster 125 and 45.92 percent confident that it is in cluster 3467. This can also be done non-linearly, where a point coming from the sum of the distances stepping one percent of the distance at a time will gain slightly more confidence than it did on a previous step. This would put an even higher amount of confidence on clusters closer to a test point. Deciding to use an exponential method or a linear method of confidence determination should be done on a case by case basis.

Case 2 In this case there is an unknown number of possible states. By changing the maximum distance allowed between points one can change the number of possible classes. Figure 2.25 shows the difference in the number of classes between setting the maximum distance to 1.5 and 2. If one knows that two modes make up the vast majority of the possible scenarios then for this example one will minimize the states by setting the maximum distance to 2. This yields two distinct clusters as well as a random unclassified point. If one tests $C = (4, 4)$ again then the distance to 3467 remains the same while the distance to 25 is now 2.91. C is now closer to 3467, but since it had previously been stated that a point needs to be no more than 2 away from another point to be in the same cluster then one can not say that C belongs to either state.

In the case where a person is forced to pick from one of the two clusters one is 50.52 percent confident that C is in cluster 3467 and 49.48 percent confident it is in

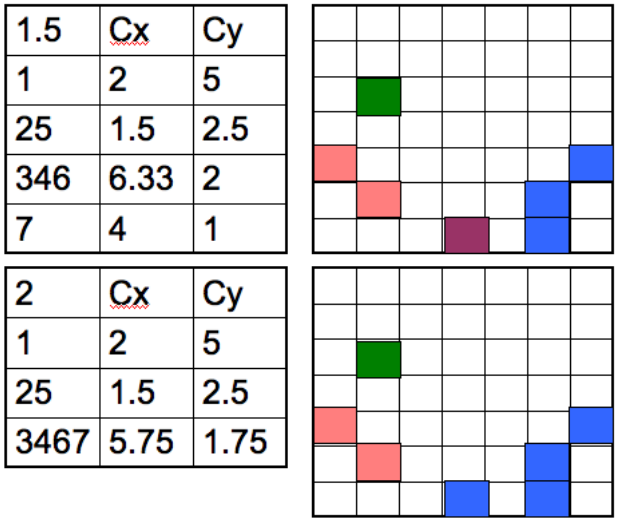


Figure 2.25: *Shows two different clustering results of the same data set with different maximum distances.*

cluster 25. If one is able to choose between each of the two clusters as well as leaving it unclassified then this can be done by introducing a third class that is exactly at the maximum clustering distance. Using three classes and the linear method of confidence calculation one now gets confidence scores of 74.22 for no classification, 63.27 for cluster 3467 and 62.5 for cluster 25. These are no longer percents because adding in more than two clusters requires an additional normalization step. Doing so yields 37.11 percent for no classification, 31.64 percent for cluster 3467, and 31.25 percent for cluster 25.

Case 3 Depending on the manner in which the clusters manifest themselves it could be difficult to separate them into two clusters. For this project if one is unable to differentiate the clusters using the distance method described above then one will need to add in more observables until some separability presents itself. When plotting the points in the high dimensional space to create the clusters one will have *a priori* knowledge of the state of each point, so the separability only needs to be present at the coordinates of the center points for each cluster.

For this example the points have been clustered a few different ways as shown

in Figure 2.26 in order to provide some insight as to how the different orientations can affect the classification of test point C. The results follow in Table 2.7.

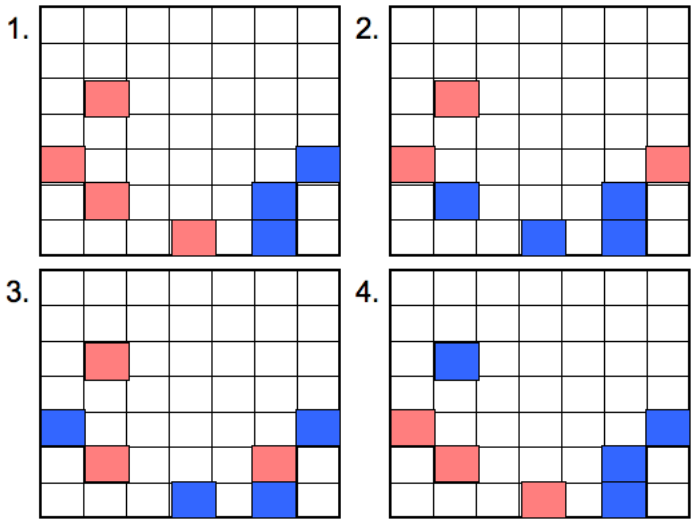


Figure 2.26: *This shows four different clustering possibilities of the same points. This demonstrates the need for accurate representations of the distribution functions of each mode in order to obtain reliable confidences.*

Table 2.7: *Distance from center of each cluster shown in Figure 2.26 to a test point* $C = (4, 4)$. *We can see that changing the manner in which each mode is described can change accuracy of the model.*

	Blue	Confidence	Red	Confidence
1	3.07	41.2	2.15	58.8
2	0.75	67.8	1.58	32.2
3	2.06	63.2	1.20	36.8
4	2.60	67.2	1.27	32.8

Case 4 In this case one can calculate confidences not only in which cluster a test point belongs but also to which sub-cluster. This is best illustrated by Figure 2.27. Calculating the confidences simply employs the math performed in the previous cases but adds an additional iteration. This will likely be completely unnecessary

in most cases as the flavors are representative of the state of different variables of data. If these variables are observed then it would make much more sense to use supervised classification.

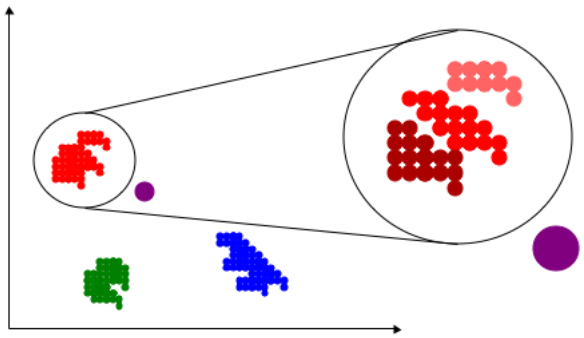


Figure 2.27: *A visual example of classifying the purple test point into one of the red, blue, or green clusters and then taking it further to attempt to classify it as one of the red sub-clusters.*

When data is missing from a vector and it cannot be filled with simulated data one can fill the hole(s) in that vector with the mean value for that observable. If all of the other observables are indicative of high flow mode and the average values move this observation closer to medium flow mode then it will become less likely to be in the anomalous high flow mode cluster. This is a good thing because with missing data there should not be as much confidence in the observation vector as there would be in one without any missing information.

Statistical Analysis

While the other methods rely on the basic descriptive statistics, a full analytic method based solely on statistics has not yet been introduced. There are two slightly more advanced methods that are worthy of exploration. One can attempt to solve for the conditional probabilities or do some regression analysis. The advantages and limitations of each are discussed below.

Conditional Probabilities The major advantage of solving directly for the conditional probabilities is that one will have the maximum likelihood probability of every state with the given inputs and available information. The rather significant downside to this method is that it calculates these probabilities by having *a priori* knowledge of the percentage of time each state occurred with each set of inputs. This is a rather unlikely scenario, however just like in past circumstances a subject matter expert (SME) might be able to assist by coming up with rough approximations of these figures. For completeness a simple example is included.

Take a very basic case where there are two possible predictor states, A and B, with two variables, each with two possible states, G and H for one, 0 and 1 for the other. This means there are 8 possible states in total, each with its own probability of occurring, as shown in Table 2.8.

Table 2.8: *An example of 2 states with 2 variables each with 2 states and the probability of each occurring.* $s + t + u + v + w + x + y + z = 1$

State	Probability
AH0	s
AH1	t
AG0	u
AG1	v
BH0	w
BH1	x
BG0	y
BG1	z

Next, the intersection of each of the two variables is calculated, simply by summing the probabilities at which both had the same value. So the intersection of H and 0 ($H \cap 0$) is simply $s + w$. Last, the probability of each of the predictor states given the state of the variables is calculated by dividing the initial probability by the result of this intersection. In other words, the probability of A given H and 0 ($P(A|H, 0)$) is equal to $s/(s + w)$. This is further illustrated in Table 2.9.

This method does not leave any room for unknown numbers of cases. That means this could only be applied to a *Case 1* type of scenario. It also has the drawback

Table 2.9: *The final calculated conditional probabilities of a simple binary state with 2 binary variable example.*

Occurrence	Probability
$P(A\|H,0)$	$s/(s+w)$
$P(A\|H,1)$	$t/(t+x)$
$P(A\|G,0)$	$u/(u+y)$
$P(A\|G,1)$	$v/(v+z)$
$P(B\|H,0)$	$w/(s+w)$
$P(B\|H,1)$	$x/(t+x)$
$P(B\|G,0)$	$y/(u+y)$
$P(B\|G,1)$	$z/(v+z)$

of requiring a finite set of values. Continuous signals can be sampled into bins in order to handle this problem. This is typically done in one of two ways: either each bin will be the same size or each bin will have the same number of members. For example, suppose one has a variable that could have any real value from 0 to 1. The two binning possibilities for 10 random values are shown in Table 2.10. The binning method chosen will be situationally dependent.

Table 2.10: *A comparison of two binning techniques on a set of 10 numbers. In the left two columns bins A and B are equal sized, with those greater than 0.5 going in B and those less than 0.5 going into A. The right two columns have bins of different sizes, splitting the data at 0.35, but both bins have the same number of members.*

Equal Bins			Equal Members	
A	B		A	B
0.1	0.6		0.1	0.4
0.2	0.6		0.2	0.6
0.2	0.6		0.2	0.6
0.2	0.9		0.2	0.6
0.3			0.3	0.9
0.4				

Regression Analysis Linear regression is a method that develops a model of the relationship between a dependent variable and one or more independent variables.

What is great about this is that it provides an equation based on each of the independent variables to yield a value for the dependent variable. Unfortunately, the dependent variable is what the analyst is trying to predict. This means that large amounts of data would have to be available in the past, or the model will have to be developed based on data from a surrogate facility. If the data has been simulated then an approximate relationship is already known and a regression is basically moot. However, one may want to mix observed and simulated data together, if they are both available.

Most statistical software packages can easily perform a regression analysis, providing the model, a goodness of fit measure (R^2), and the amount of error. In a single variable regression R^2 is simply the square of the correlation coefficient. In a multiple regression scenario it is the ratio of the sum of squared error in the model to the sum of squared error from simply using the mean. During the course of this project MATLAB was used to perform ordinary least squares regression, as was SPSS, which has a regression button with many flashy options that go beyond the scope of this discussion. A full description of regression in SPSS is available in *Discovering Statistics Using SPSS (and sex and drugs and Rock 'n' Roll)* by Andy Field.

One of the more useful applications of regression is the use of stepwise regression. Given a list of independent variables used to predict a dependent variable, the independent variables are only inserted into the model if they contain useful information. What this means is that if two variables describe essentially the same amount of variability in the dependent variable, only one of them will be used in the generation of a regression model (Field, 2009).

2.4.2 Weighting

One may want to use different weights on the variables, depending on their origin. For example, simulated variables should not be treated the same as real variables, and one may not want to put as much weight on a high order principal component as is placed on a lower order one. A confidence level may be applied in these cases

to represent the quality of the data. Quantifying these in a precise manner is nearly impossible, however one can usually rely on various quantitative methods to make a good decision as to what the confidence levels should be. For example, given just the information that the tanks are typically at a level of 4-8 inactive, a simulated variable of this data should not be used at all. Not until this information is tied with the way flow changes with time of day can one put any faith in this variable, perhaps using the correlation of the two variables in some way to determine the weight. Once more information becomes available, such as how tanks adjust with rain and season, the confidence level should increase.

Weighting variables is very easy to do in template matching; it simply changes the number of variables from an integer quantity into a floating point quantity (like 6 to 5.5). In geometric analysis the weighting is applied by decreasing the range of values upon normalization. Since a regular variable is given values from 0 to 1, a half weighted variable would only have values from 0 to 0.5. Regression analysis determines the weights on all variables: real, simulated, high order, low order.

2.5 Summary

In this chapter several different pieces of a method that can be used to identify the operational mode of an industrial facility using multi-modal data were outlined. The final step will be an interpretation of the results of the predictive algorithms. Clearly these will be site, mode, signal, and situation dependent, but one will have to determine if the results warrant any actions. If all signs point to nefarious activity then it is clear that some action needs to take place. However, if all possible modes have the same probability then something is wrong. Either data is missing and a new type of sensor needs to be used to aid in the process, there is not enough separability amongst the observables, or perhaps there is too little confidence in the data that has been used for the analysis. Either way, new requests will need to be made, and the new data will have to be acquired and ingested into the system so it can be laid into the site model for analysis and the process can be run again. While

this may sound inconclusive and possibly frustrating, if done properly it can lead to the appropriate and supportable conclusion that sufficient data is not yet available to direct a course of action, which is still a very important conclusion and can help to prevent situations such as the war in Iraq.

This chapter is summarized in Figure 2.28. The next chapter discusses the site and goes over the data that was collected over the course of this project.

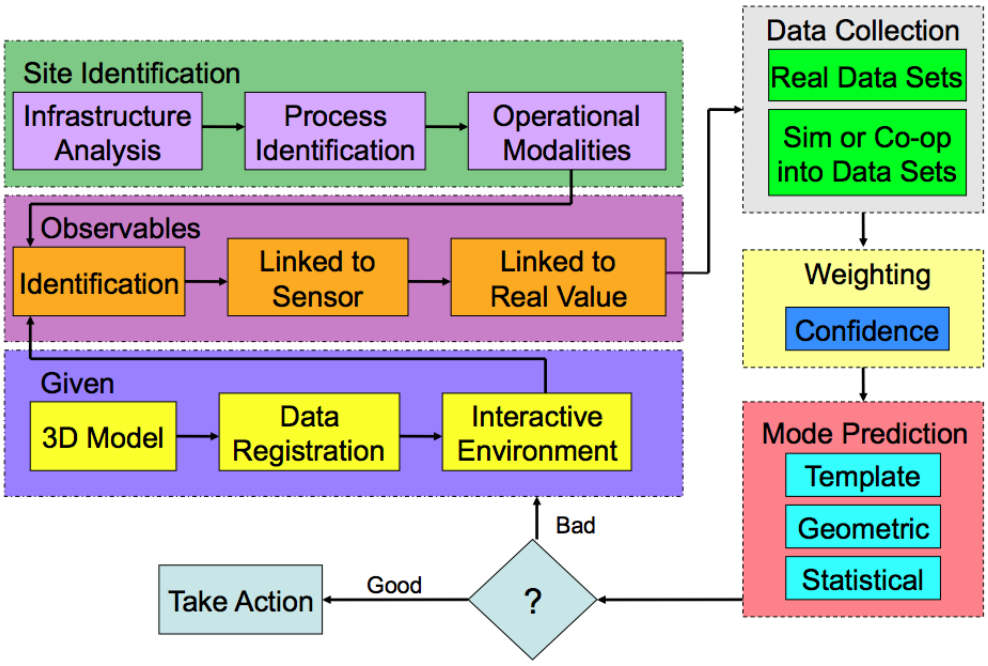


Figure 2.28: *This diagram shows the operational mode identification process broken down into six key parts. The blue section (yellow boxes) lists the pieces already present that are needed for this project. The green section (purple boxes) shows the three key pieces of site identification. Those first two sections feed into the purple section (orange boxes) where the observable identification process is shown. From there it goes on to the data collection process (grey section, green boxes). This follows on into the yellow section (blue boxes) where the different signals will receive a weighting based on the confidence an analyst associated with it. The last red section (aqua boxes) is where different algorithms are utilized to predict the operational mode. If the results from this are good then prudent action can be taken. Otherwise it will be necessary to re-examine the data and determine what can be done to improve the results.*

Chapter 3

Site of Interest: Data and Processes

The goal of this project is to develop a method for identifying processes at industrial facilities. This cannot be done through typical remote sensing practices. In order to do this one has to delve deeper than image information and acquire knowledge about the facility being studied. This will typically require the use of an outside expert as well as data from non-image sources. The difficult task then becomes quantitatively interpreting the multi-modal data.

In this chapter the first few steps from the method shown previously in Figure 2.28 are applied. Again, given that 3D data registration exists and there is a repository in which to place the data, the steps that follow are site identification, determining the observable signals associated with this site, followed by data collection.

3.1 Van Lare Site

Imagine you are an analyst and you have just been tasked with figuring out everything there is to know about a facility. All you are given is the location of the site and a few images. What do you do next? This is exactly how this project began.

For the exploratory work it is desirable to have a constrained and local target (i.e. facility) so a site with appropriate characteristics needs to be identified. The test site to be investigated for this research is the Frank E. Van Lare Wastewater Treatment facility in Rochester, NY. Obviously a nefarious site of real interest is unlikely to be local. An analyst would have to go through the procedure outlined in Section 2.2 in order to identify the facility. This is an exhaustive process, requiring lots of research and expensive data collects that may or may not yield good results. In addition, the resources to do this are not available to academic departments.

This site was chosen in large part for its proximity to RIT, but there are several other factors that made it a good candidate. It is on the small side as far as industrial facilities go, and the local government granted us limited access. Most importantly, the Imaging Science department at RIT has been collecting data of Van Lare for over a decade because of its location relative to other sites of interest to the department, such as the Genesee river plume. This data repository provides the project with a small amount of temporal data that can be used to build a knowledge base of the facility.

The other important reason for choosing this site is that there is nothing sensitive going on within the facility, making the information accessible and distributable to people working outside of the project. There are several complex processes taking place within that can be observed and modeled.

The primary function of the Van Lare facility is to treat wastewater and release it back into nature. In completing this function some solid and gaseous waste products are made so there are some supporting processes on the site as well.

3.1.1 Wastewater Treatment

Van Lare uses the activated sludge process for wastewater treatment. This is a common method that has several variations at the various wastewater treatment facilities around the world. Van Lare implements the activated sludge process in four phases: grit removal, aeration and primary settling, secondary settling, and chemical treatment.

Wastewater Intake

Wastewater is brought into the site from a state of the art, complex, underground city sewer system (shown in Figure 3.1) as well as being pumped from many of the surrounding suburban areas. The intake from the city has three forms: storm drains, sewer drains, and combined. All of the wastewater and other objects that end up in the pipes flow to the plant predominantly using gravity, because the plant is downhill from the city. There are some pump stations, however, that are located at various locations throughout the greater Rochester area because the coverage is widespread and it is not downhill from everywhere. The surrounding suburban areas have shut down their wastewater treatment facilities and are pumping their wastewater to VanLare, the largest of which is Irondequoit. Irondequoit has its own elaborate tunnel system shown in Figure 3.2. To handle this task there are several different pumps, but the main pump house shown in Figure 3.3 contains seven pumps powered by very large motors. Next to this pump station can be seen a small transformer yard. One can see changes in the transformer's relative retrieved signal in the thermal band. This change is dependent upon the number of pumps that are running at collection time.

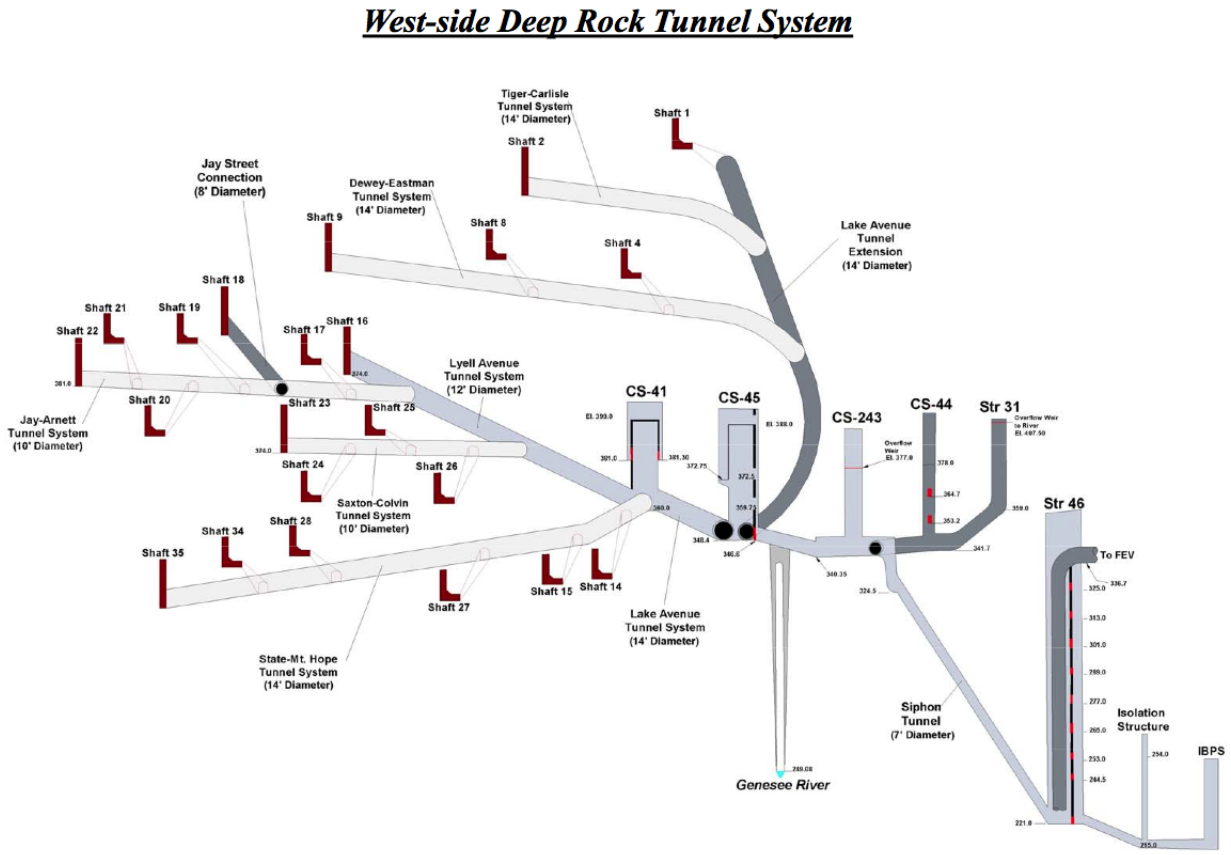


Figure 3.1: *The deep rock tunnel system for storing wastewater underneath the city of Rochester. Image courtesy of Monroe County.*

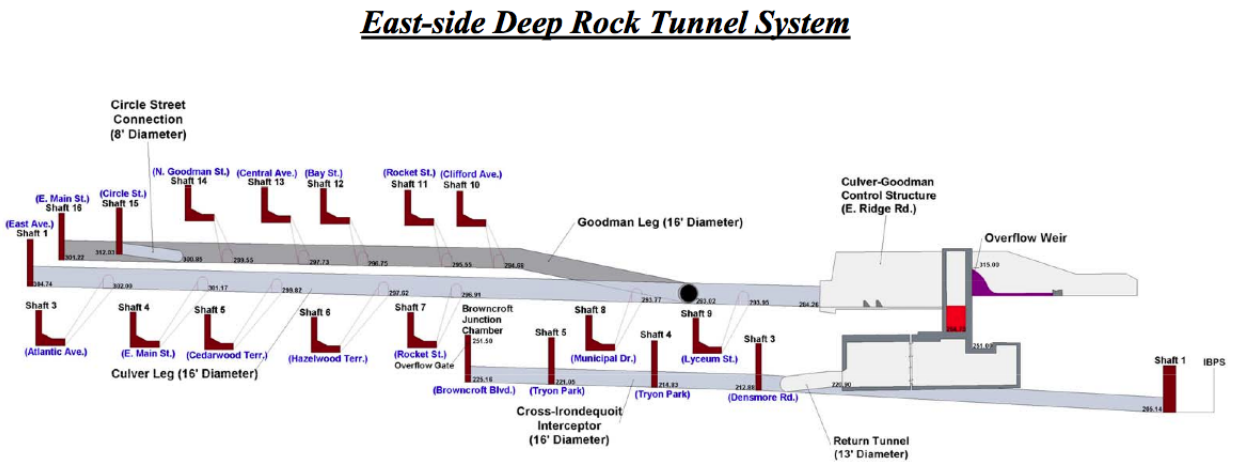


Figure 3.2: *The deep rock tunnel system for storing wastewater underneath Irondequoit. Image courtesy of Monroe County.*

Figure 3.3: *The pump station and transformer yard that bring wastewater to the Van Lare facility from Irondequoit.*

Phase 1: Grit Removal

The first part of the separation process occurs in the buildings shown in Figures 3.4 and 3.5. People flush all sorts of things down the toilet and other objects are washed out of the street and into the storm drains. There are four main processes that are done at any wastewater treatment plant. Grit removal is used to eliminate the majority of solid waste products such as rocks and garbage. The plant has two sides each running slightly different versions of the same processes. Along the west side of the plant the wastewater goes through a screen that removes the solid materials from the water and pulls them into a garbage bin. The water then travels into a chamber that uses air to cause the water to move in a cyclonic fashion. This causes things like sand and small stones that go through the screens to settle to the bottom and get pulled out into the garbage. The east side of the plant performs this process in a slightly different manner; the wastewater still goes through a screen but in a separate building, then undergoes grit removal through a flow control system. Solid objects within the wastewater sink to the bottom of a tank and are scraped away.

Figure 3.4: *West side screening and grit removal building, which uses large moving screens to grab the objects and pull them out of the liquid and into a garbage bin, then spins the wastewater to cause grit to settle to the bottom.*

(a) East side screening. (b) East side grit removal.

Figure 3.5: *The east side screening and grit removal process which takes place in two large buildings. The screens grab large objects out and discard them as trash and the grit removal building slows the flow down to allow grit to settle to the bottom.*

In order to help illustrate what happens to the wastewater as it goes through the plant, some simple cartoons have been provided. In the phase one cartoon, shown in Figure 3.6, one can see that dirty wastewater goes through the grit removal process, and comes through still dirty, but with all large chunks of garbage having been removed.

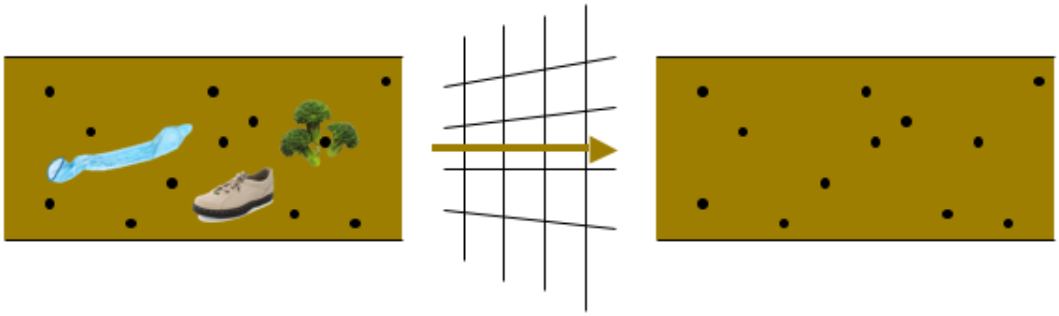


Figure 3.6: *Wastewater as it enters the plant and goes through the grit removal process.*

Phase 2: Aeration and Primary Settling

Aeration and settling happen next, with wastewater often making several trips back and forth between these two processes. The large arrays of tanks, shown in Figures

3.7, 3.8, and 3.9, are used to provide the right amount of oxygen to the naturally occurring bacteria in the wastewater so that it can effectively break down the organic material into floc. The floc then clump together and come out of suspension and settle to the bottom (Bartlett, 2007).

The wastewater is aerated in a bottom up approach in the object seen in Figure 3.7. This is a more modern method of aeration, and there are still a large number of older top down aeration tanks shown in Figure 3.8. In addition to providing the air with a nice wastewater mist, these tanks create what is called a mixed liquor, which contains the aforementioned biomass. This discharge is pumped to the primary settling tanks.

Figure 3.7: *West side aeration tanks, which are covered because the air is added to the mixture through pipes at the bottom of the tank.*

Figure 3.8: *East side aeration tanks that vigorously stir the wastewater.*

On the east side there are three large circular tanks (Figure 3.9(b)) and on the west side there is a series of smaller rectangular tanks (Figure 3.9(a)). With the correct amount of aeration the flocculation process can take place, allowing clumps to settle to the bottom of the settling tanks in a substance called sludge (Bartlett, 2007). Most of this sludge undergoes some separate processes that will be discussed later, but some of it is recycled back into the aeration tanks. The microorganisms are in a delicate balance that needs to be maintained in order for the process to function properly; recycling some sludge helps to keep this balance. The cleanest of the wastewater skims from the top and is about 95 percent pure as it moves on to the third phase.

(a) West side (b) East side

Figure 3.9: *Settling tanks in which the flocculation process occurs, allowing most organic material to be removed from the wastewater. The west side uses several small rectangular tanks while the newer east side tanks are much larger and circular.*

Looking at the illustration in Figure 3.10 one can see the wastewater go through a bottom up aeration process. This wastewater then goes to the primary settling tanks where the heavier solid material sinks to the bottom and the lighter, cleaner wastewater is skimmed from the top to move on to the next phase.

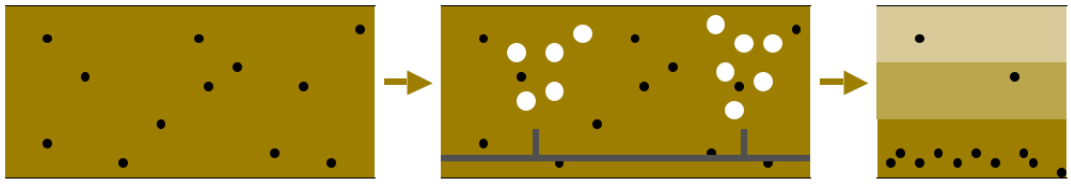


Figure 3.10: *Wastewater as it receives aeration and then enters the primary settling process.*

Phase 3: Secondary Settling

The wastewater from the two sets of settling tanks flows into a common set of secondary settling tanks shown in Figure 3.11. Similar to the last tanks, the wastewater is stirred slowly, allowing higher density materials to sink to the bottom as sludge. The wastewater skims from the top and flows on to the last phase. At this point the water is about 99 percent pure, but still unsafe to drink (Lukas, 2007).

Figure 3.11: *The six large secondary settling tanks.*

As is evident from a visual inspection of the tanks in Figure 3.11 compared to the tanks in Figure 3.9, the wastewater is significantly more clear in the secondary process. This is illustrated in Figure 3.12. What this cartoon is attempting to show is that despite the wastewater being predominantly clear at the top layer, there is still organic material present that is treated in the fourth phase.

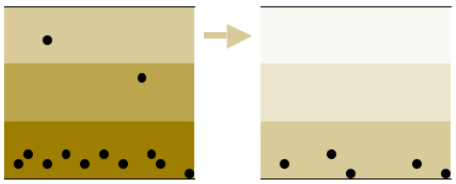


Figure 3.12: *Wastewater as it goes through the secondary settling process. Notice that while the top layer is mostly clear, it is still not ready to be released into the lake.*

Phase 4: Chemical Treatment

There are disease causing microorganisms present in the water that are not handled by the activated sludge process. The goal then becomes to kill everything in the water so that it can be safely discharged. A chlorine mixture similar to that used in swimming pools is added to the water, and it is sent on a 15 minute trip through a mazing contact tank, shown in Figure 3.13. The water is then pumped out 3 miles into the middle of Lake Ontario, during which time the chlorine levels in the water are supposed to have decreased to a non threatening level (Monroe, 2010).

Figure 3.13: *A mazing tank used to mix chlorine in with the wastewater to kill off harmful microorganism not handled by the activated sludge process.*

Supporting Processes

From the bottoms of the primary and secondary settling tanks most of sludge is diverted to one of the circular structures shown in Figure 3.14 (remember that

some of the sludge is recycled to maintain the biological balance in the aeration and primary settling tanks). These structures are sometimes called evaporators, or thickeners, but are essentially settling tanks where the sludge can sit for a long time and dry out. They are covered because the odor released from these tanks is strong and unpleasant. Some of the pipes that can be seen are carrying air on to a tertiary process described later in this section.

Figure 3.14: *Evaporators or thickeners, these large structures are long term settling tanks for sludge.*

Settled sludge goes to the area shown in Figure 3.15. The white building is full of very large centrifuges that spin the sludge to separate out even more water. These are loaded, turned on, spun, then emptied - not a continuous process like most of the other areas of the plant. Once the cycle is completed, the sludge still contains a very small amount of water and "has the consistency of carrot cake" (Lukas, 2007). Still, this is as dry as the sludge gets at the plant, so it goes into the back part of the building with the green roof. Here it is placed in giant hoppers where it awaits pickup by Waste Management, which uses the dried sludge as filler in its land fill.

The sludge contains all of the organic material that was previously in the wastewater now in a concentrated form. The odor this produces is difficult to handle, to say the least. The air that comes in contact with this sludge is pumped into the area of the building that is adjacent to the smoke stack. Inside the air is scrubbed and purified using coal filters in order to eliminate as much of the odor as possible and to keep from disturbing plant employees as well as residents that live nearby (Lukas,

Figure 3.15: *The group of buildings that deal with sludge treatment.*

2007). There are still complaints from time to time that the odor is escaping the plant, which is indicative of either too much air being forced through the coal filters at once, or that the current filter needs to be replaced.

The smoke stack on the back of the building does not fit with any of the current processes, causing it to spark some interest. According to Monroe County the plant used to burn all of the sludge after it was done drying, to further decrease its volume. There used to be two smoke stacks connected to their own incinerators to handle all of the waste. Incineration stopped in 2005 when an arrangement was made with Waste Management to use the solid material in a landfill. Incineration is still available as a backup process but, according to plant employees, has not been used (Monroe, 2010; Lukas, 2007).

An analyst checking out an unfamiliar site would not simply read one article and assume this to be true. To verify the dormancy of the smoke stack one would require a form of persistent surveillance. Fortunately, with the stack being so much taller than the other buildings at the plant, one could easily monitor it from off the site. A simple camera would not do, since many harmful gases are transparent in the visible spectrum. This would likely require a thermal camera staring at the smoke stack for long periods of time (in the event of a detected gas, a hyperspectral sensor would be needed to aide in determining the type of gas). Getting a sensor

set up is simple enough, but if this was a nefarious site it is unlikely one would be able to stare at the stack for weeks on end without detection. This would require several different collects at separate times and locations. Ultimately believing that the stack is in fact dormant would depend on several things. Obviously none of the collects would be able to yield any positive results, but it would also rely on other signals from the plant. Suppose the plant is using significantly more electricity than needed, or various other signals manifest themselves that go outside the normal operating procedures. These would lead one to believe that something odd is going on, and perhaps the data collections thus far have just been unlucky.

A Tale of Two Sites

Based on the information provided previously, one can deduce that the Van Lare plant behaves like two separate plants. There are two different forms of grit removal, aeration, and primary settling tanks. This is because of a massive undertaking by the city of Rochester and Monroe county in the 80s and into the 90s. That is when the construction of the tunneling systems, shown in Section 3.1.1, took place. This tunnel system was made to direct more wastewater to the Van Lare plant, so it had to be expanded. The plant has been around since 1916, so the technological advances that have taken place have caused there to be significant differences between the original design and the more recent addition.

Envisioned as part of the AANEE project is the 3D environment in which to immerse an analyst in all of the data. All data includes all different data types, as well as all pieces of historical data. An analyst should be able to select a modality and scroll through time, all measurements of that type in a given location. For example, using even some very low resolution image data from LANDSAT of the Van Lare facility, one can see some major changes. In Figure 3.16 there are two images of the Van Lare site, one from 1980 and the other from 1990. Clearly from these two images one can see that the east side operation was constructed during this time.

Figure 3.16: *These are zoomed in LANDSAT images of the Van Lare site. The middle image is from 1980 and one can clearly see only a few grey pixels. The right image is from 1990. The same grey region of pixels is there, but the arrows are pointing to two distinct new features, which are the newer sets of primary and secondary settling tanks. LANDSAT images courtesy of the USGS.*

3.1.2 Van Lare Process Model

While much of this information was obtained by taking a tour of the facility, the vast majority of this information could have easily been obtained through the use of a few airborne images and a subject matter expert (Note: from my experience with Van Lare I am able to look at other wastewater facilities and get a good idea of the flow of materials there, and I am certainly not an expert). An analyst can now map all of the above information to an image of the site. In Figure 3.17 there is an airborne image of the site and, after labeling everything that was previously mentioned, there are two buildings left without a label. These two buildings should now become a priority in the site investigation. In order for a site to truly be doing what it claims there should not be any missing pieces to the puzzle nor should there be any extra pieces of equipment.

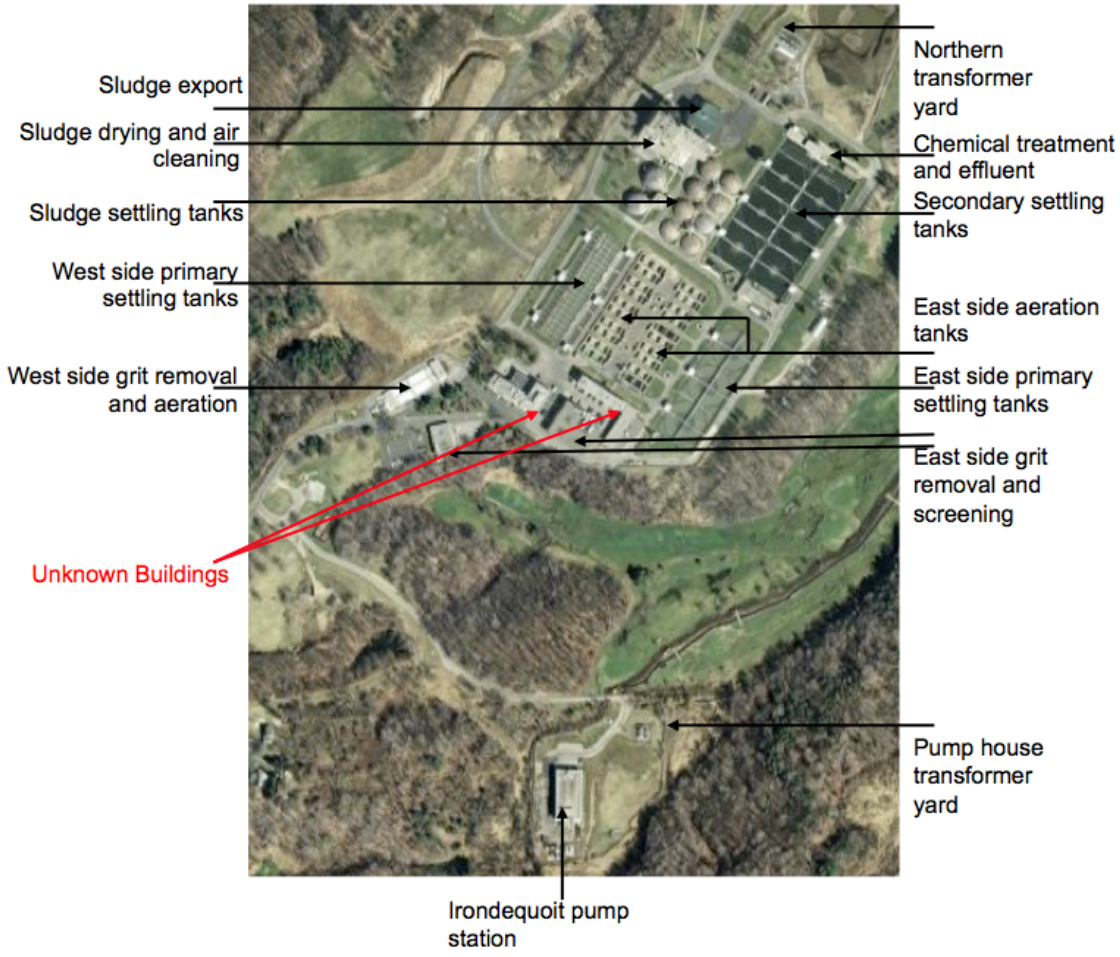


Figure 3.17: *Van Lare Wastewater Treatment Plant as seen in Google Earth.*

Imagery and the subject matter expert were able to easily identify one of the two buildings. The building shown in Figure 3.18 has many windows and is located near a parking lot. This is typical of many office buildings, meaning this was the administration building. There are also several small chimneys on the roof. Wastewater treatment plants have to test their water regularly to make sure the biological and chemical contents are balanced appropriately. These chimneys are there so that odors and vapors resulting from the testing may easily escape the building.

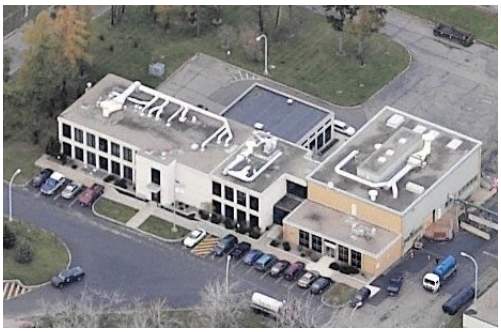


Figure 3.18: *One of the unknown buildings from Figure 3.17, this building is a combination of administrative offices and wastewater testing facilities. Image courtesy of Bing Maps.*

The function of the other building, shown in Figure 3.19, was easily discovered through online research on the Monroe County website to be a maintenance building. One wing is dedicated to electrical maintenance projects while the other wing is more for mechanical maintenance. Verifying this, however, would be significantly more difficult. This building is located in the middle of the plant, making it hard to monitor from ground locations around the perimeter. Persistent surveillance of the building would be difficult (though not impossible) from the air, but might be necessary to catch the employees walking in with broken equipment and leaving with fixed or new pieces.

One can now combine the four phases used to complete the activated sludge form of wastewater treatment at this plant and build a basic flow model. It then becomes possible to overlay this model on the plant to obtain an improved perception of the processes described above as demonstrated in Figure 3.20. The arrows represent

Figure 3.19: *One of the unknown buildings from Figure 3.17, this building is used for maintenance projects. Image courtesy of Bing Maps.*

the different materials as they make their way through the plant during normal operation.

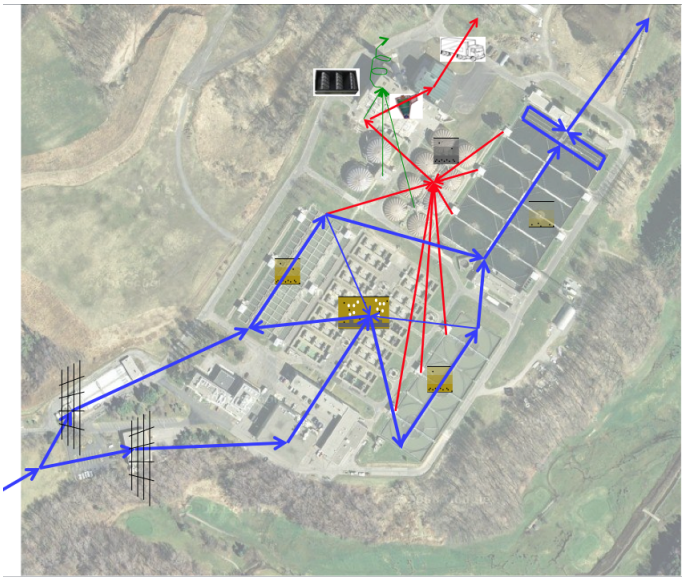


Figure 3.20: *Wastewater treatment process overlaid on an image of the plant. Blue arrows are wastewater, red arrows are sludge, and the green arrows are air. Image courtesy of Google Earth.*

One can take this a step further and integrate the process model into the Advanced Analyst Exploitation Environment software. Figure 3.21 shows a primitive model of the wastewater as it makes it way through the plant. This is useful for analysts as they work to determine the relationships amongst all of the site infrastructure. One can imagine a more advanced systems that shows the wastewater and supporting processes for all different operation modes. A visual could have lines that increase in thickness to represent different levels of flow, arrows that change color based on the clarity of the water, and infrastructure that talks when clicked to explain its purpose in the process.

Figure 3.21: *Process model integrated into the AANEE software. The dots move from left to right showing the flow of wastewater through the plant.*

These flow diagrams are the foundation of a sites operations and are a very important step in the operational modality identification process. Without a thorough understanding of the underlying processes taking place at a site it will not be possible to proceed to the next step. Developing this knowledge of Van Lare and wastewater treatment in general was a major portion of this project. It is important to note that the vast majority of this knowledge came through interviews with subject matter experts, with some support through internet research. Multiple tours of the Van Lare facility were taken, but the knowledge base on the activated sludge process was started at the small treatment plant in Cazenovia, NY. Van Bartlett, the head

trainer at SUNY Morrisville and the teacher for some of the SMEs utilized at Van Lare, was by far the most valuable asset on this subject. Learning the process and seeing it applied in different ways and scales would be a benefit to any analyst when tackling a problem of this magnitude.

3.1.3 Operational Modalities

The different operational modalities of a site that an analyst seeks to identify are going to be site dependent. Upon investigation of the Van Lare site it is evident that there are two sides of the plant running similar operations. At some plants it could happen that one side is running while the other is off and they regularly switch and one might want to be able to determine which side is active at a particular point in time. At the Van Lare site, however, both sides are running concurrently almost all of the time so it is unlikely that we would detect that over the course of this project (Lukas, 2007). It does, however, make for an interesting alternative operation scenario and will be discussed in Section 4.5.

It is also possible that a site with multiple processes can only have a subset of those processes running at one time. For example, if the Van Lare site transfered all of its wastewater into the aeration tanks, executed that process, then emptied them into settling tanks to run that process, and neither was active at the same time. An analyst may want to be able to determine how many subsets there are and which subset is active at a particular point in time. Again, at the Van Lare site all of the processes are running continuously so that type of analysis does not apply here.

Van Lare's operation can be broken into the following operational states:

1. Normal - the regular day to day operation of the facility
2. Shut Down - when the plant has been completely turned off
3. Bypass - during instances of high rain the storm water goes through grit removal, can be given chemical treatment, and then heads out to the lake
4. Single Side - when only one side of the plant is functioning

The plant has not shut down in over 10 years, nor has there been a significant emergency, thus making it difficult to detect these particular modes. However it is possible to conceive of the site being shut down. There would be no activity in any of the tanks, no motors running, and the transformers would all have the same temperature as the air. The bypass is completely underground so its detection is nearly impossible. We can, however, guess that the plant is running in bypass mode during any major rain event in the spring, when the water table is at its highest. Normal mode is the easiest to detect, which is what the plant has been in every day for the last 10 years.

Normal operational mode can be broken down into three states: low, medium, and high. Given hourly flow rates from July 1, 2007 - June 30, 2008 (which will be discussed more later) medium flow has been defined as the mean of these values +/- one standard deviation. Using this method suggests that it should be in medium flow approximately 68% of the time. The data is not Gaussian, however, and the plant is actually in medium flow mode 85% of the time. While it would be more convenient to differentiate between modes with equal probabilities, it is unlikely that any facility would have such a scenario. For example, most of the time a nuclear reprocessing facility is in standard reprocessing mode and hopefully only rarely is it in nefarious mode.

Single side operation refers to one of two possibilities. As mentioned before this site is indicative of two different plants due to major updates that were performed in the 1980s. One option for single side mode would be that only one set of the processes are running, either east side or west side. The other option for single side mode would be that all wastewater is coming from one source, either from the city of Rochester or from Irondequoit.

3.2 Information Collection

Data collection is a major step in understanding a site of interest. Collecting overhead images, walking around the outside of a site taking pictures, and getting tours

of surrogate sites will help an analyst get an idea as to what it is one is looking at. But in order to take advantage of all information available on a site of interest one needs to go beyond ordinary data collection methods. It is also necessary to understand the relationships that different pieces of infrastructure have with one another on a quantitative level. To do so one must enlist the help of a subject matter expert (SME). These are people that could run the site of interest if needed. They understand the processes taking place at each piece of infrastructure and can help an analyst fill in gaps in their data by developing some preliminary guesses as to the correlation of different signals as well as their probability distribution functions. This section will discuss and show the raw data that has been collected from Van Lare, and then step through the process by which discussions with an SME can turn it into simulated quantitative data.

3.2.1 Data Collection

When looking at an object for the first time, analyst or not, people will want to see the object first in the environment in which they are most familiar. When it comes to a remotely sensed site of interest, that environment is typically a true color image. Fortunately, thanks to Google Earth and other free global imaging resources available online, it is very easy to do that. Doing so allows one to become familiar with all of the visible infrastructure. A sensor should be tasked based on what is being looked for (additional details can be found in Chapter 4). Since the goal of this project is to examine many scenarios, several forms of collects were performed.

Overhead Imagery

This project uses seven multispectral data collects, six of which were taken during the day, and one at night, collected using the Wildfire Airborne Sensor Program (WASP) sensor that is owned and operated by the Digital Imaging and Remote Sensing (DIRS) group at RIT. WASP has six bands on four cameras: red, green, and blue bands on a Terrapix camera, and three Phoenix infrared cameras. The Phoenix-Near camera is sensitive in the 0.9μm to 1.7μm region (SWIR) with a band

center at 1.3μm. The Phoenix-Mid camera is sensitive in the 3μm to 5μm region (MWIR) with a band center at 4μm. The Phoenix-Long camera is sensitive in the 8μm to 9.2μm region (LWIR) with a band center at 8.6μm. A typical atmospheric transmission spectrum is shown in Figure 3.22 (Schott, 2007). The day time collects were on January 18, 2007; July 24, 2007; August 1, 2007; May 3, 2008; May 13, 2009; and June 29, 2009. The nighttime collect was from May 13, 2009. Other data has been collected by Pictometry, Digital Globe, Google Earth, AVIRIS, and several other sensors. While this data will be used in this study the main focus will be on the data collected from the WASP sensor.

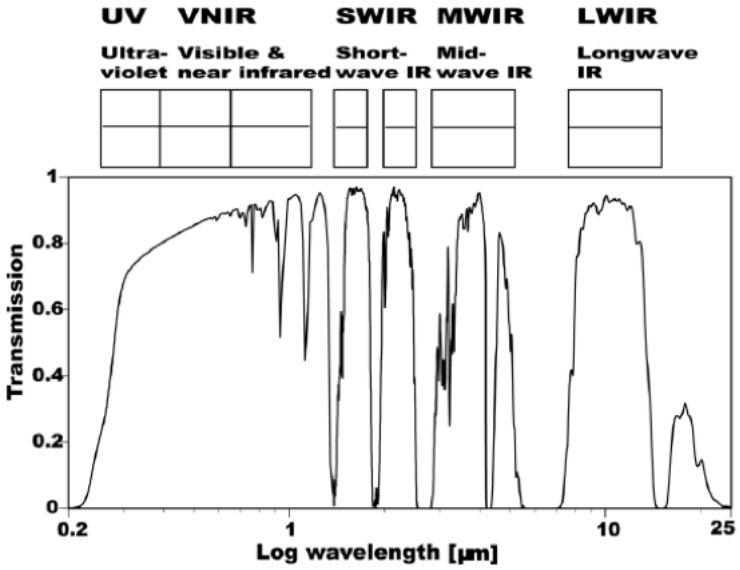
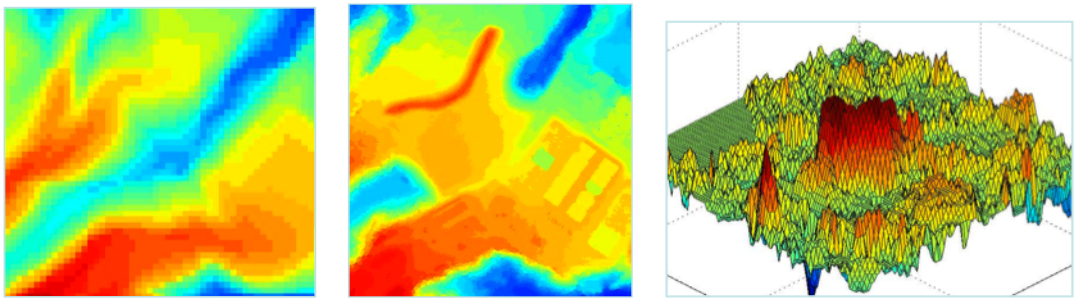


Figure 3.22: *A diagram of the spectral regions that traverse through the atmosphere and are utilized in remote sensing. Image courtesy of Schott, 2007.*

With overhead imagery being rather abundant there should be a database that automatically collects these images and runs several target and change detection algorithms. For example, shortly after completing a LIDAR collect the data was compared to the digital elevation model (DEM) that was already in place for Van Lare. The old model had 30 meter resolution while the new model is 1 meter resolution - a vast improvement. That amendment, however, did not account for

what was found when a change detection was performed. As is shown in Figure 3.23 there is a large area where the elevation increased dramatically. This is believed to be where some of the removed dirt was placed in order to put in the aforementioned underground sewer system.



(a) The original 30 meter DEM (b) The newer 1 meter DEM (c) The difference in elevation

Figure 3.23: *Elevation models of Van Lare several years apart. Blue is low, red is high, and one can see that a valley has been filed in, evidenced further by the large red hump in the right-most image. Images courtesy of Karl Walli.*

On Site Measurements

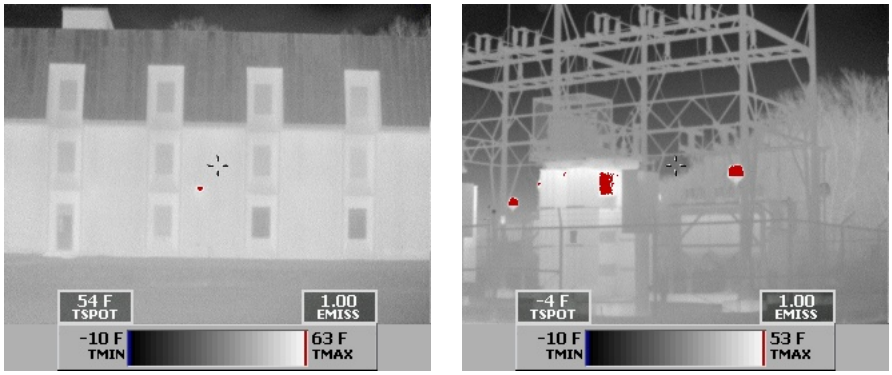
Since the Van Lare site is a cooperative facility it was possible to get onto the site on multiple occasions and take measurements. The initial visit was a tour of the facility during which time several pictures were taken with a Nikon D50, a few of which are shown in Figure 3.24. From this it was possible to get a good grasp on the flow of the material through the plant, and a basic idea of the purpose for which each individual machine was used.

A walk around the plant in the middle of the night with a thermal camera was also performed. Most of the images captured during this visit were useful only in that they showed there was very little activity. All of the buildings on the site looked similar to Figure 3.25(a), where there is clearly nothing interesting to observe. However, Figure 3.25(b) is very interesting. One can see that the transformers are vastly warmer than the objects around them. This brought about further research,

(a) Settling tanks (b) Pipes connecting buildings (c) Underground pumps

Figure 3.24: *Some of the pictures from the first trip to Van Lare.*

and lead to the discovery that these transformers have a direct relationship to the giant pumps bringing wastewater from Irondequoit.



(a) Typical building at night (b) Transformer yard at night

Figure 3.25: *Ground based LWIR images of Van Lare.*

On one occasion all of the vehicles were counted as they drove on and off the site for an entire business day. This was done from a spot just inside the gate and did bring about some trouble. It was requested that this not be done ever again. The number of sludge removal trucks, the number of septic trucks, and the total number of other vehicles were all counted and are summed up in Table 3.1. The amount of traffic was significantly higher than originally anticipated. We were under the impression that this was just a wastewater treatment plant and that there were no other activities taking place on the grounds. On the contrary, Van Lare functions as a parking lot for county vehicles. Several county employees travel to the plant

just to pick up a van or pick-up truck and then head off to run some errand. For this reason counting the number of vehicles in the parking lot during the day will not provide insight as to the number of employees on site. This information could still be used for information about night shifts.

Table 3.1: *Vehicular traffic at Van Lare, 5/13/2009*

Date	Sludge Trucks	Septic Trucks	Total Vehicles
5/13/2009	12	7	290

Despite being told never to count vehicles again on site, it was possible to return just outside of the gate where very little attention would be paid. This time a camera with a fish-eye lens was used in order to experiment with the test postulated in Section 2.3.1. The camera was set up across from the gate where a few different experiments were performed. With such a wide angle lens one can see vehicles as they drive across for an average of 11.3 seconds. The ones that slow down and drive into the site combined with the ones that drive out from the site are viewable for an average of 20.2 seconds: 9.1 seconds in front and 11.1 seconds on plant property. During rush hour approximately ten percent of the vehicles that drive by enter the plant, while the rest of the business day approximately sixty percent of the vehicles that drive by enter or exit the plant. After business hours it becomes significantly easier because in order for someone to enter or exit the plant one must wait for a gate to open. Since often an analyst is presented with too much data a few different sampling rates were tested, the results of which are shown in Table 3.2.

Table 3.2: *Table of sampling interval, vehicle detections, and amount of data.*

Sampling Interval	Detections	Images Per Hour
1 second	Perfect	3600
5 seconds	Perfect	720
10 seconds	Perfect	360
15 seconds	Near Perfect	240
30 seconds	Vast Majority	120

Based on the summary made available in Table 3.2 one can have perfect vehicular

detection at a sampling interval of 10 seconds, which would get 2-3 images of each vehicle entering and exiting the site. This would only use a tenth of the bandwidth of the 1 second case, which would make it much more realistic to get real time results. In the case where it is not necessary to have perfect detections, as in with Van Lare, a lower sampling interval can reduce the amount of data one has to transmit and process and the overall results can simply be scaled. For example, taking an image every 30 seconds will get every truck that enters and exits the plant (they are significantly slower than cars and pick-ups) but will certainly miss a small percentage of the small vehicles going in and out. This can simply be compensated for by knowing ahead of time that one will miss approximately ten percent of the vehicles so one can multiply the detected results by 1.11 to obtain the estimated total vehicular traffic.

If bandwidth, storage, or overabundance of data is a serious issue one can cut down the amount of data even further by utilizing a motion sensor during the low traffic times during the middle of the night. While most industrial facilities are operational twenty-four hours a day the amount of vehicular traffic in and out of a facility as well as passers-by often decrease during the late hours of the night into the early hours of the morning. At Van Lare an entire hour can pass without any traffic. During these times data collection can be cut significantly by activating a sensor that captures an image every time any movement or sound is detected.

The pump house was made accessible on one occasion to see the different types of motors they had. There are two Teco-Westinghouse three phase induction motors with 700 horse power at 710 rotations per minute (RPM), a Teco-Westinghouse with 1250 horsepower at 507 RPM, and four General Electric reliance induction motors with 1500 horse power and 592 RPM. It is potentially possible to differentiate between which pumps are running based on the RF emanations.

Mass Media Intelligence

The Internet has a few interesting stories about Van Lare. The sewer system in Rochester is one of the most advanced in the world at preventing overflow problems

(Rochester History, 2010). From news stories it was possible to find out that Van Lare no longer burns any of their sludge. The one smoke tower that is left is no longer in use and the sludge instead gets exported to the Waste Management landfill (Monroe Country, 2010).

From the public records it was possible to request hourly flow information from June 1, 2007 - June 30, 2008. This data was combined with the hourly rainfall data provided by the National Weather Service. From this data it was able to perform a series of statistical analyses. The first step was to determine the operational modes based on flow rates. This was done by taking the average flow rate and defining everything within one standard deviation of the mean as being medium, while everything below that was low and everything above that was high. Of the 9504 points, 555 are low (5.8%), 7875 are medium (82.9%), and 1074 are high (11.3%).

The influent flows to the plant are a combination of sewage and storm water. When the levels of storm water are too high that water gets redirected. The large objects are removed and it bypasses the aeration and settling phases to receive chemical treatment. It is then combined with the effluent flow (Lukas, 2007). Doing so obviously has an impact on the quality of the effluent water from the plant. Table 3.3 shows that there is a positive correlation amongst each of the influent lines and the amount of precipitation. During this time period it was unusually dry to the point of the summer of 2007 being a drought. This causes a decreased correlation amongst the amount of rain and the pumps because a large portion of the rain was absorbed into the ground. One can also notice that there is a rather significant difference in the correlations 6 hours of rain has with the two influent sources. This is likely due to the different amounts of time it takes for rainwater to reach the plant from the two sources and the greater distance Irondequoit is from the Rochester airport, where the rainwater measurements are made.

Table 3.3: *A correlation matrix of the two influent pumps, the storm system siphon, and the amount of rain over the previous 6 hours from June 1, 2007 - May 31, 2008.*

	Rochester	Irondequoit	Storm Drains	Rainfall
Rochester	1	0.4817	0.3614	0.2679
Irondequoit	0.4817	1	0.5096	0.3435
Storm Drains	0.3614	0.5096	1	0.5002
Rainfall	0.2679	0.3435	0.5002	1

3.2.2 Information Collection

As mentioned before, in order for an analyst to utilize every resource it is necessary to go beyond ordinary data collection. This section outlines the process by which qualitative information about a variable can be turned into simulated quantitative data, following the guidelines and estimates laid out by a subject matter expert. There will be more examples of this in Chapter 4.

SME Interrogation

When first talking to a subject matter expert about a site it is not always clear to know which questions to ask. It is best to begin by letting the SME explain all that is known about the process and how it is applied to a familiar location. Giving an analyst a tour of the facility is very helpful to gain insight about the goings on inside of buildings. One will want to know each phase of the main process and all of the supplementary processes taking place, and it is useful to take pictures of everything so as to have a visual reminder of each step. Next one will want to have the SME explain how these processes are applied at the new site. It is likely a good idea to get multiple experts to do this individually so as to have a high confidence in the information. For this project two experts were used, as well as a few other random wastewater employees for questions pertaining to their specific area of expertise on the plant.

Ask Questions

Knowing what each piece of equipment does and why is fairly easy, but knowing how each piece relates to each other is difficult. After taking some tours (whether real or virtual) an analyst should certainly be able to come up with several questions. At Van Lare, as well as most wastewater treatment plants, the objects that stand out the most are open water tanks. One clearly might want to know the relationship these objects have with one another. With no data being recorded as to how many tanks are being utilized at any particular time, one is simply left with the option of speaking with a SME. Upon inquiring about the settling and aeration tanks, it was discovered that the number of tanks in use is "highly correlated" with the amount of wastewater going through the plant. Further, the number of aeration tanks being used is typically close to the same as the number of settling tanks being used, often differing by one or two. And last, there are typically 4-8 inactive tanks of each type (Lukas, 2007).

Relate Information to Collected Data

Initially this may not sound too helpful, but combined with the data previously collected, this is actually a wealth of information. It is already known that the plant is typically running in a normal flow mode. Further, it is known that it is running at a medium flow over 80% of the time. If there are typically 4-8 inactive tanks of each type, then it is logical to assume that these 4-8 tanks are inactive during medium flow mode. One can also assume that when all twenty of them are inactive that there is no flow, and when none of them are inactive the plant is at the high end of high flow. There are 838 points that are high flow, so as an initial estimate it is assumed that the plant has 0 inactive tanks when it is in the top 25% of those occurrences. Putting all of this together one can develop the series of points shown in Table 3.4.

From there one can simply generate a graph and get a best fit line. In this case the best fit was a polynomial of order 2. This has the advantage of generating an equation that directly relates the number of inactive tanks to the flow. This graph

Table 3.4: *A collection of estimated data points based on SME information. The SME stated that there are typically 4-8 inactive tanks, which I took as indicative of medium flow mode. If all of the tanks are inactive then there is no flow, and it is assumed that if the plant is in the top 25% of high flow mode (189.4 mgd) then there are 0 inactive tanks. Flow is measured in millions of gallons per day (mgd).*

Flow(mgd)	Inactive Tanks
189.4	0
149.8	4
110.2	6
70.6	8
0	20

is shown in Figure 3.26.

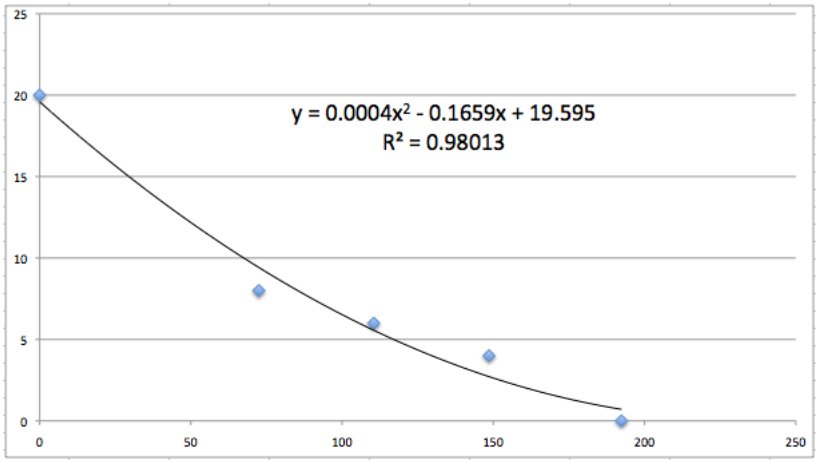


Figure 3.26: *A graph showing tanks vs. flow generated from the points in Table 3.4*

Generate Simulated Data

The ultimate goal is to have a way to approximate real data when it is not entirely available. Real data rarely follows the perfect curve of the equation provided in Figure 3.26 and very rarely has a correlation coefficient (the square root of the R^2 value on the graph) greater than 0.9. In order to simulate more realistic data one must introduce small amounts of variability when generating the data. One can do

this by adding white noise - a random variable with zero mean and a small range - to the original simulated values.

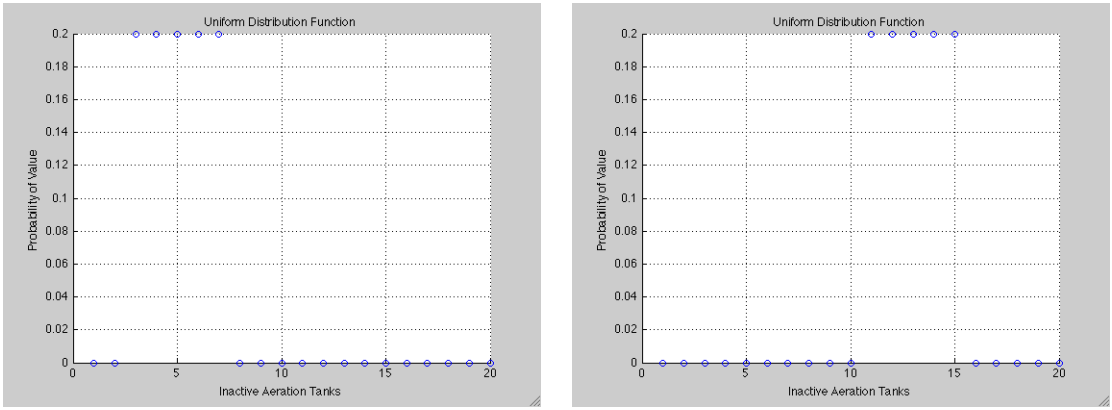
As an example here is a demonstration of how one could go about simulating data for the settling and aeration tanks. To begin one can use the equation provided in Figure 3.26, using the 9504 data points of flow information to generate the number of inactive settling tanks. Start with something simple, like a possible 25% swing from the equation to introduce some variability. Since the SME mentioned in Section 3.2.2, the different types of tanks often differ in the number being used at any one time by one or two tanks one can generate a random integer from -2 to 2, inclusive, and add it to the number of inactive settling tanks to get the number of aeration tanks, making sure that neither one could exceed twenty or drop below zero. Because of cases of extremely high flow some adjustments will likely have to be taken into consideration. Since two variables being highly correlated often refers to a correlation coefficient of 0.7 or higher, one will not want to drop below that for either type of tanks when compared to the flow (Field, 2009). Running the simulation in this manner should yield results similar to those found in Table 3.5.

Table 3.5: *A correlation matrix of inactive settling and aeration tanks along with wastewater flow through the plant.*

	Inactive Settling Tanks	Inactive Aeration Tanks	Flow
Inactive Settling	1.000	0.8962	-0.8306
Inactive Aeration	0.8962	1.000	-0.7379
Flow	-0.8306	-0.7379	1.000

Adjusting the aeration tanks in this manner is called implementing a *latent model*. A latent model is simply a statistical model that is used to relate one set of observed variables to a set of variables that was not observed but are understood and can be modeled. These are sometimes called inferred variables (Bishop, 1999). In this case it is assumed that the inactive settling tank variable is observed and the inactive aeration tanks are inferred, and a discrete uniform distribution function is then applied. This is demonstrated in Figure 3.27. If a more advanced scheme is desired it would be possible to introduce a Gaussian curve that represents the

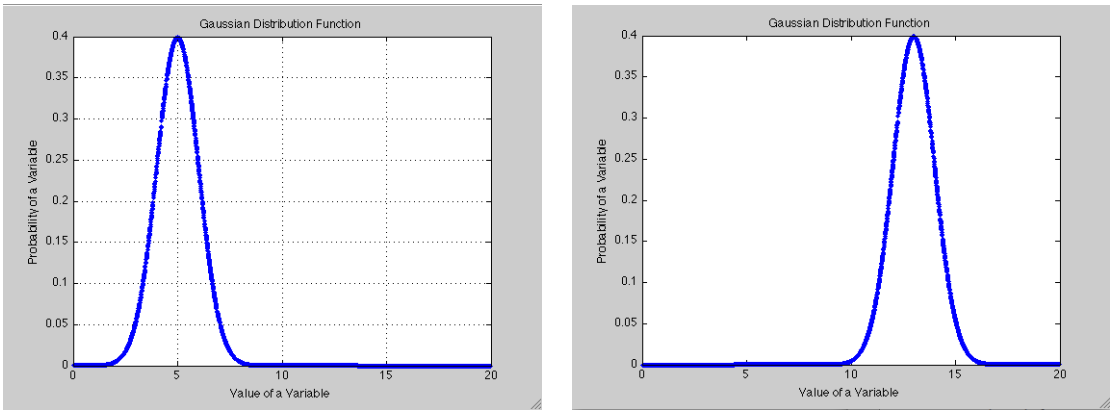
probability of all possible values. The shape of the curve remains fixed, representing a constant standard deviation, but the peak of the curve moves based upon the value of the input variable. While the number of inactive tanks is a discrete variable this is not the case for all variables. A continuous example is shown in Figure 3.28.



(a) Input Value = 5 (b) Input Value = 13

Figure 3.27: *A discrete uniform probability function based on the input of another variable. In the first image, the input value from the observed variable is 5. In the second image the input value is 13, causing the data points to shift to the right.*

Of course, if an analyst is trying to predict flow levels then simulating data based on flow is not allowed in most situations. It would be possible to do this if one were using data from a cooperative site, with modulations made based on the differences in the two sites explained by an SME. Still, there are ways around this as will be demonstrated in Section 4.1. The important thing to note here is that given enough information from a SME, approximations for missing variables can be made to aid in the prediction of plant activities. Given enough data over time one will likely be able to improve upon the accuracy of these approximations, however it will still be important to distinguish between real data and simulated data when making the predictions, because simulated data should bring about increased uncertainty in the analysis.



(a) Input Value = 5 (b) Input Value = 13

Figure 3.28: *A continuous probability function based on the input of another variable. In the first image, the input value from the observed variable is 5. In the second image the input value is 13, causing the curve to shift to the right. The shape of the curve remains unchanged.*

3.3 Summary

At this point an intelligence analyst knows significantly more than desired about wastewater treatment, but this knowledge has helped to paint a vivid picture of all of Van Lare. This chapter also covered an example of the process of observable generation. This is a very important step in this project as data simulation is a key component to process identification. An analyst must be able to rely on SME information to produce accurate models and generate simulated data in order to fully take advantage of all available information. An example of data simulation was used to demonstrate potential utility, which will become even more evident in the following chapter. The discussion continues using different methods for quantitative analysis of the data on different scenarios for the Van Lare facility.

Chapter 4

Van Lare Mode Prediction

With the data in place and many of the subprocess relationships known it is possible to approach the analysis stage of this process through the use of various scenarios. The metrics described previously in Section 2.4.1 have been used on a few different potential scenarios of the Van Lare facility. These follow from the list given in Section 3.1.3. All of these test scenarios use the AANEE concept of having a large amount of registered data described in the previous chapter. For each scenario there will be a discussion of the observable signals that are relevant to the test, the sensors that are used to gather this data, the data that was used to run these tests (real and simulated), and the results of the tests. In the event of simulated data, the process by which the data was generated based on the SME input will be shown.

4.1 Flow Prediction

Since the Van Lare facility is almost always operating under normal conditions attempting to predict the amount of wastewater going through the plant at a particular point in time is an obvious choice for the first test of the method described in this thesis. In performing this particular test the model relating tanks to flow illustrated in Section 3.2.2 will have to be obtained differently, since one does not want to simulate tank information from the variable that is being predicted. Instead one should

begin with the hourly rain data since there is hourly flow data and hourly rain data for the same periods.

4.1.1 The Data

The rain variable was expanded into several variables to show cumulative rain for 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 14, 16, 18, 20, 22, 24, 36, 48, 60, and 72 hours. The correlation of each rain variable to the flow is shown in Table 4.1. If one wants to use a single variable to represent the rain it is best to pick the one that has the highest correlation with flow. In this case that is at 12 hours, and that will yield an R^2 value of 0.1687. However, there are other methods that may yield even better performance.

Table 4.1: *The amount of correlation amongst rain and wastewater flow.*

1 hour	2 hours	3 hours	4 hours	5 hours	6 hours	7 hours
0.296	0.336	0.361	0.378	0.391	0.400	0.405
8 hours	9 hours	10 hours	12 hours	14 hours	16 hours	18 hours
0.408	0.409	0.410	0.411	0.408	0.404	0.399
20 hours	22 hours	24 hours	36 hours	48 hours	60 hours	72 hours
0.393	0.386	0.379	0.344	0.332	0.323	0.313

One way to better utilize the rain information is to use the principal components. Principal components are orthogonal vectors that are ordered by the dimension of most variability. The principal components, the cumulative explained variance in the rain, and the R^2 value achieved when performing regression analysis to predict flow are shown in Table 4.2. The table shows that a better result can be obtained by using the first two principal components, but very little improvement is seen after adding in the third.

Looking at the coefficients of each of the principal components one can sometimes see which portions of the data comprise each component. This is easiest when all of the coefficients have the same sign and admittedly gets confusing when they do not. The coefficients for the first five rain principal components are shown in Figure 4.3.

Table 4.2: *The number of principal components of rain used, how much variance in the rain they cumulatively explain, and the R^2 value of a regression model when these PCs are used to predict the wastewater flow at Van Lare.*

Number of PCs used	Cumulative Variance Explained	R^2
1	80.7	0.1457
2	93.8	0.1736
3	96.67	0.1946
4	98.19	0.1959
5	98.83	0.1971

The first principal component is clearly dominated by the 36-72 hour cumulative rain amounts, the second has most its information in the 12-24 hour cumulative rain amounts, and the third has the largest contributions from the 7-10 hour regions (and a large contribution from 72). The shorter cumulative rain amount, 1-4 hours, only play a significant role in the tenth principal component, which explains very little of the variance. This says that overall short term rainfall has very little effect on the flow at Van Lare.

Table 4.3: *A table showing the coefficients of the first five principal components of the rain data.*

	PC1	PC2	PC3	PC4	PC5
1	0.0078	0.0188	0.0332	-0.0417	0.0414
2	0.0164	0.0397	0.07	-0.087	0.0845
3	0.0255	0.0618	0.1069	-0.1302	0.1216
4	0.035	0.0842	0.1415	-0.1676	0.1467
5	0.0447	0.1065	0.1721	-0.1964	0.1573
6	0.0547	0.1283	0.1973	-0.2141	0.1515
7	0.0647	0.1493	0.2161	-0.2191	0.1302
8	0.0749	0.1691	0.2275	-0.2111	0.0943
9	0.0851	0.1875	0.2311	-0.1909	0.0471
10	0.0953	0.2043	0.227	-0.1594	-0.0069
12	0.1157	0.233	0.199	-0.0723	-0.1139
14	0.1359	0.2551	0.1499	0.0296	-0.1915
16	0.156	0.2705	0.085	0.1289	-0.2162
18	0.1756	0.2785	0.01	0.2106	-0.1825
20	0.1946	0.2792	-0.0692	0.2634	-0.0954
22	0.2129	0.2726	-0.1456	0.2805	0.0233
24	0.2306	0.2589	-0.215	0.2618	0.1457
36	0.3259	0.0962	-0.4864	-0.1739	0.5352
48	0.406	-0.1154	-0.355	-0.4809	-0.1977
60	0.467	-0.3168	0.0472	-0.146	-0.5091
72	0.5042	-0.4684	0.4291	0.3746	0.3658

This analysis can be continued in order to find out what affect the time of year has on the flow of wastewater. In the spring when the snow melts there is likely to

be more water in the storm drains and therefore higher flow levels than there are in the winter when much of the precipitation remains frozen on the ground. Also, there is going to be a connection between the hour of the day and the amount of flow because at night there are more people sleeping than there are awake. Most people have a daily routine that involves sleeping through the night hours, not flushing toilets or drinking water nearly as often as we do during the day. This time of day dependence was calculated by finding the mean flow for each hour of the day. The results are shown in Tables 4.4 and 4.5.

Table 4.4: *Probability distribution function for flow per month.*

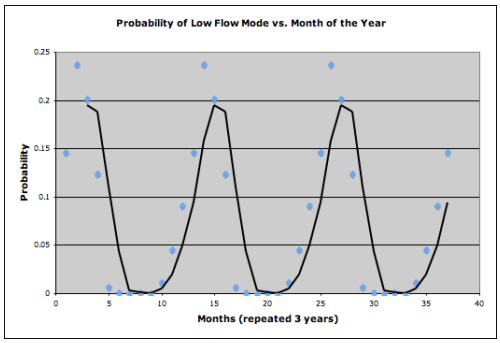
Month	P(Low)	P(Medium)	P(High)
January	0.0	96.1	3.8
February	0.0	78.0	22.0
March	0.0	61.3	38.6
April	0.0	84.0	16.0
May	1.2	96.9	1.9
June	4.6	91.2	3.6
July	10.4	88.4	1.2
August	14.6	85.4	0.0
September	23.8	72.5	3.8
October	20.2	76.8	3.1
November	12.4	80.3	7.4
December	0.7	84.3	15.1

The probabilities in all of these figures jump around a little and with several years worth of data it would be possible to smooth them out. In that situation one could use piecewise linear interpolations to generate curves. An alternative way to smooth them out would be to generate approximating functions. There are several different methods available. A few that are listed in *Introduction to Numerical Analysis Using MATLAB* by Rizwan Butt were implemented, but the results were not vastly different than those provided by curve fitting software. Using the 'Add Trendline' function in Excel using second order polynomial one can generate the curves shown in Figures 4.1 and 4.2. This program gives the option of adding a

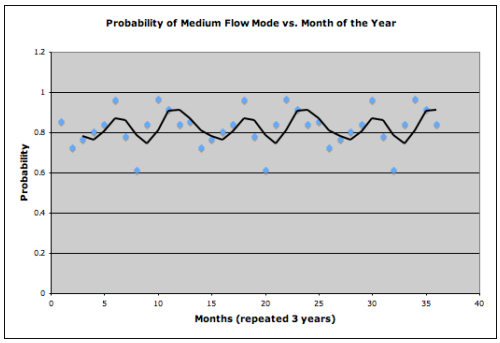
Table 4.5: *Probability distribution function for flow per hour.*

Hour	P(Low)	P(Medium)	P(High)
0	2.3	85.3	12.4
1	6.2	83.4	10.4
2	11.4	79.2	9.4
3	16.0	75.9	8.1
4	18.9	73.9	7.2
5	20.2	71.7	8.1
6	22.5	70.4	7.2
7	21.4	70.5	8.1
8	14.0	77.9	8.1
9	9.7	82.1	8.1
10	5.2	86.7	8.1
11	2.9	85.4	11.7
12	1.0	87.3	11.7
13	0.6	88.0	11.4
14	0.3	87.0	12.7
15	0.0	87.0	13.0
16	0.0	87.3	12.7
17	0.0	84.1	15.9
18	0.3	84.7	14.9
19	0.0	85.1	14.9
20	0.0	85.4	14.6
21	0.0	87.0	13.0
22	0.3	87.7	12.0
23	0.3	87.3	12.3

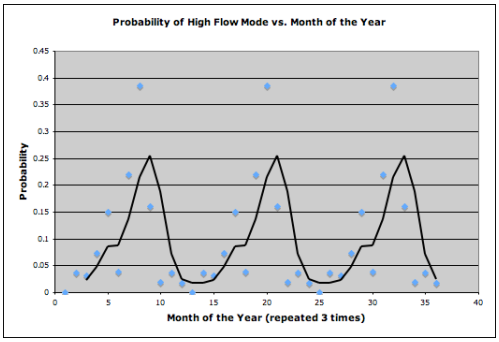
line, logarithmic curve, polynomial curve, exponential curve, or moving average to fit the data. Expanding the data to repeat and show the cycle will only yield good results when using the moving average. The approximate equations represented by each graph are easily calculated. The monthly ones are shown in Equations 4.1, 4.2, and 4.3, while the dailies are shown in Equations 4.4, 4.5, and 4.6.



(a) Low



(b) Medium



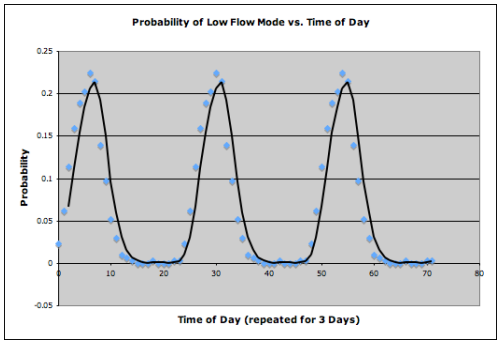
(c) High

Figure 4.1: *The probability distribution curves generated in Excel from the data points in Table 4.4. Given a day of the month one can plug that value into the given equations and get an approximate probability of each flow mode.*

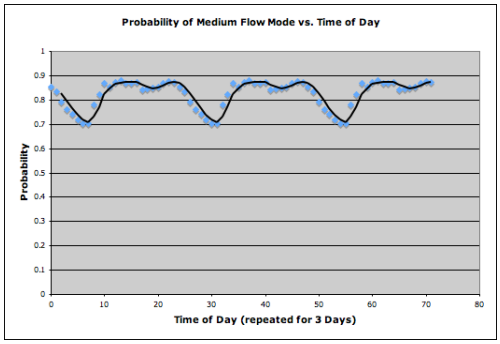
$$P(Low|Month) \approx \frac{sin(\frac{\pi x}{6})}{10} + 0.1 \tag{4.1}$$

$$P(Medium|Month) \approx \frac{cos(\frac{\pi x}{3})}{10} + 0.8 \tag{4.2}$$

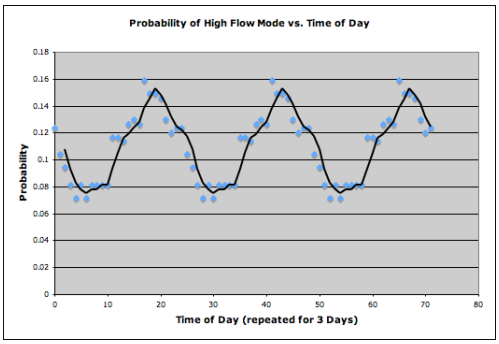
$$P(High|Month) \approx \frac{sin(\frac{\pi(x-4)}{6})}{10} + 0.15 \tag{4.3}$$



(a) Low



(b) Medium



(c) High

Figure 4.2: *The probability distribution curves generated in Excel from the data points in Table 4.5. Given a time of day one can plug that value into the given equations and get an approximate probability of each flow mode.*

$$P(Low|\,Time\ of\ Day) \approx \frac{sin(\frac{\pi x}{12})}{10} + 0.1 \tag{4.4}$$

$$P(Medium|\,Time\ of\ Day) \approx \frac{sin(\frac{\pi(x-12)}{12})}{10} + 0.8 \tag{4.5}$$

$$P(High|\,Time\ of\ Day) \approx \frac{sin(\frac{\pi(x-12)}{12})}{25} + 0.12 \tag{4.6}$$

Another signal to be used would be the amount of inactive aeration or settling tanks. If the number of tanks being used at a particular time is known, then one

can simply use the inverse of the process demonstrated in Section 3.2.2. Instead of generating a function that predicts the number of tanks based on flow, generate a function that predicts the flow based on the number of tanks. Using the same points that were used in Table 3.4 except reversing the axes, one can obtain the graph and equation shown in Figure 4.3.

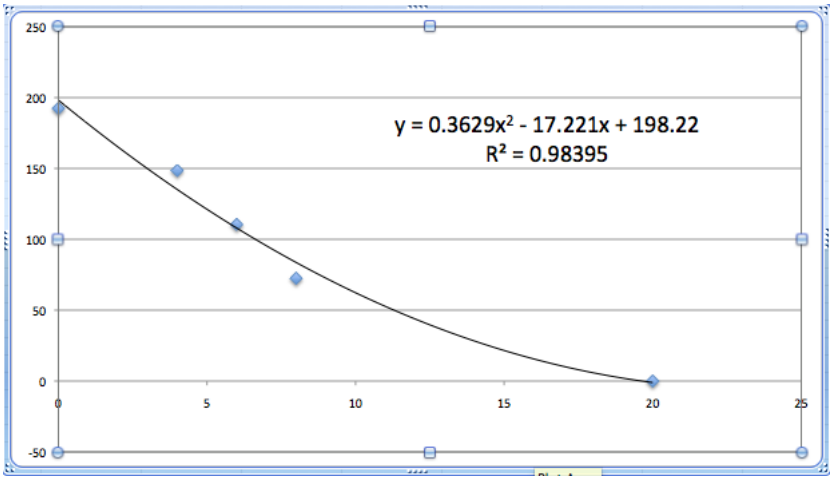


Figure 4.3: *A graph and function that predict the flow based on the number of inactive tanks.*

If the tank data is not available then it could be simulated. However, it might be of interest to simulate the data in a different manner than was shown as an example in Section 3.2.2 because flow is what is being predicted. Instead, since the relationships between rain and time with flow are known, one can generate simulated tank data based on these two variables. This will obviously be less accurate, but it is additional information that can aid in the prediction. In order to do this as accurately as possible the steps performed in Section 3.2.2 will need to be repeated. First, inquire of the SME as to the effect rain has on the tanks. It was discovered that the tanks are set to be active a few hours before a rain event, depending on the severity of the incoming storm. They will also remain active for several hours after the rain event, again depending on how much rain fell (Lukas, 2007).

This simulation is a tad complicated. Begin with the information that there are typically 4-8 inactive tanks, and link this relationship with the daily changes in flow.

In the early hours of the morning when low flow mode is more likely, make those 8 inactive settling tanks. And in the evenings when high flow mode is more likely, make those 4 inactive tanks. Add in a random sliding scale to complete the cycle. It should be very likely to increase going from 7 a.m. to 5 p.m., as is demonstrated in Figure 4.4. Likewise, it should have a probability to decrease from 5 p.m. to 7 a.m. as is demonstrated in Figure 4.5. This is the implementation of the aforementioned latent models. The range of tank values available at each point changes based on other input, namely the time of day and the value of tanks on the previous instance.

Last look at the rain information and look to see how long a storm lasts and how much rain is falling per hour. The more total rain, the fewer inactive settling tanks for longer periods of time. This algorithm is shown in Figure 4.8. To test the quality of the simulated data see how well it correlates with the flow data. While it is unlikely to achieve the 0.7 standard of high correlation, it was possible to achieve 0.31, which is a good result with simulated data based on the information available. Since these variables are available for any point in time one can use them to predict the number of inactive tanks at any given time. Note this process was used to simulate inactive tank data. In practice the data would come from image analysis.

Next make adjustments based on the time of year. Randomly add 0-2 in the late summer and fall and subtract 0-2 in the spring (shown in Figures 4.6 and 4.7).

There are several other types of data that can be helpful but were not all thoroughly tested. These are signals which at least one SME provided some information about the relationship with flow. To get quantitative values for these variables one must use the process outlined in Section 3.2.2.

Irondequoit Pump House Begin with transformer yard outside the pump house. Two questions worth asking are: "Are these transformers related to the pumps?" and "In what way does the temperature of the transformer relate to the amount of power being drawn?" Since the SME was an expert on wastewater treatment, the answer to the second question was unfortunately, "I have no idea," but he was confident when he said that the transformers are in fact there to help power the pumps (Lukas, 2007). From what little that was learned about transformers the temperature will

$\text{InactiveTanks}_{700} = 8$
$\text{InactiveTanks}_{1700} = 4$
FOR j = 8 TO 16
    x=randu(1)
    IF x > 0.7
        InactiveTanks_j=$\text{InactiveTanks}_{(j-1)}$ - 1
    ELSEIF x < 0.05
        InactiveTanks_j=$\text{InactiveTanks}_{(j-1)}$ + 1
    ELSE
        InactiveTanks_j=$\text{InactiveTanks}_{(j-1)}$
    END IF
    IF $\text{InactiveTanks}_j > 20$
        $\text{InactiveTanks}_j = 20$
    ELSEIF $\text{InactiveTanks}_j < 0$
        $\text{InactiveTanks}_j = 0$
    END IF
END FOR LOOP

Figure 4.4: *The algorithm used to get a random slide of inactive tanks from 8 a.m. to 4 p.m.*

```
FOR j = 18 TO 6
        x=randu(1)
        IF x > 0.1
                InactiveTanks_j=InactiveTanks_(j-1) + 1
        ELSEIF x < 0.01
                InactiveTanks_j=InactiveTanks_(j-1) - 1
        ELSE
                InactiveTanks_j=InactiveTanks_(j-1)
        END IF
        IF InactiveTanks_j > 20
                InactiveTanks_j = 20
        ELSEIF InactiveTanks_j < 0
                InactiveTanks_j = 0
        END IF
END FOR LOOP
```

Figure 4.5: *The algorithm used to get a random slide of inactive tanks from 6 p.m. to 6 a.m.*

```
x=randu(1)
IF x < 0.5
        InactiveTanks_j=InactiveTanks_j + 1
ELSEIF x < 0.9
        InactiveTanks_j=InactiveTanks_j + 2
END IF
IF InactiveTanks_j > 20
        InactiveTanks_j = 20
END IF
```

Figure 4.6: *The algorithm used to add a random amount of inactive tanks to the simulated late summer and fall values.*

```
x=randu(1)
IF x < 0.5
        InactiveTanks_j=InactiveTanks_j - 1
ELSEIF x < 0.9
        InactiveTanks_j=InactiveTanks_j - 2
END IF
IF InactiveTanks_j < 0
        InactiveTanks_j = 0
END IF
```

Figure 4.7: *The algorithm used to add a random amount of inactive tanks to the simulated spring values.*

```
IF Rain_j > 0
        count = 1
        k = j
        WHILE Rain_k > 0
                k = k + 1
                count = count +1
        END WHILE
        tanks = round(count / 3)
        InactiveTanks_j = InactiveTanks_j - tanks
        IF count>4
                InactiveTanks_{j-1} = InactiveTanks_{j-1} - tanks
                InactiveTanks_{j-2} = InactiveTanks_{j-2} - tanks
                InactiveTanks_{k+1} = InactiveTanks_{k+1} - tanks
                InactiveTanks_{k+2} = InactiveTanks_{k+2} - tanks
        END IF END IF
IF InactiveTanks_j < 0
        InactiveTanks_j = 0
END IF
```

Figure 4.8: *The algorithm used to add a random amount of inactive tanks to the simulated spring values.*

in fact change based on the amount of power being drawn (Copper, 2010). This still does not really tell one too much about how this relates to flow so it is necessary to ask more questions. Going back and forth one will eventually learn that about 30% of the flow is typically coming from Irondequoit, and if all of the pumps were to be running at once it is possible to have the plant running at high flow levels (Lukas, 2007).

In order to know what temperature would be reached by the transformers, if they were running at capacity, one would have to know the temperature rise rating. This is something that is associated with the highest temperature the transformer can achieve (Federal Pacific, Summer 2011). The power company was not cooperative in sharing this information, and it is unlikely one would be able to easily obtain this information from a non-friendly site. Instead one can rely on the aforementioned process described in Section 2.3.2 and use the ratio of the LWIR signal over the transformers to the LWIR signal around them. It was assumed that the three data collects for which flow information is known were running at that typical 30% level from the pumps. From the ratios given in Table 2.5, one can link 1.026 with a flow level of 26.97, 1.035 with a flow level of 25.31, and 1.042 with a flow level of 32.68, 30% of the flow that was going through the plant during those collects. A ratio of 1 will be achieved when there is no flow going through the pumps, and based on the temperature rise ratings of various transformer types, one can say that the maximum ratio of 1.15 is achieved with a flow level of 148.68 mgd (millions of gallons per day). This then generates the linear expression shown in Equation 4.7.

$$TransformerRatio = 0.001 \cdot Flow + 1.0045 \tag{4.7}$$

This formula can now be used to generate hourly data to go along with the flow data, adding in some white noise to make it more realistic. It is also potentially interesting to be able to generate data which simulates different percentages of flow going through the pump house. This can be done by simply changing the percentage of flow level inserted into the equation. Obviously, with more knowledge of the transformers, more data collects with corresponding flow information, and a better

model of heat dissipation from transformers much more accuracy would be possible. For example, in what manner does rain affect the temperature? It will clearly cool it down at least a little, but how much rain would be needed to make them appear to be off? These are exactly the types of things the AANEE environment would help to bring together in order to produce more accurate models.

An alternative way to collect pump information would be to place passive RF sensors in the area around the pump house. Initially the signals would be used to determine how many different pumps were running. The signals detected would change whenever a pump was turned on or off, and with time it would be possible to determine how often the pumps operate and to what capacity. The pumps are deep underground so a seismograph may also be helpful.

When asked about how the number of pumps running changes with flow, the SME was not able to provide a straightforward answer. This is because not all of the pumps are the same size so they can handle different amounts. They are run on a rotating basis, to prevent overheating and overuse. For low levels of flow one of the smaller pumps could be running, and for medium levels of flow there could also still be one large pump running. In general, if only one pump is running it is most likely during low flow time and if more than four pumps are running it is most likely during high flow times (Lukas, 2007). Since high flow mode is anything above 148.68 mgd, one can mark that amount as using 4 pumps and also link a really low flow mode (50 mgd) as using 0 pumps. This yields the linear relationship in Equation 4.8.

$$NumberOfPumps = round(0.04 \cdot Flow - 2.03) \tag{4.8}$$

Centrifuges The centrifuges in the sludge handling portion of the facility are large machines that require big motors. Like the pump house, RF sensors can be used in the area around the centrifuges.

Through SME interrogation it was possible to discover several things about these machines. As mentioned previously in Section 3.1.1, the centrifuges do not run a continuous process. There are four of them with typically only one running at a

time. They are used to further dry sludge that has been separated from wastewater that arrived, on average, about 48 hours earlier. They spin the sludge for one hour, separating out as much water as possible. The RF sensors would be able to detect the number of centrifuges running at a time.

Given that the average flow is 110.189 mgd, it was assumed that one centrifuge would be running almost continuously at this level. That means that as soon as one is done spinning, another starts. With lower levels of flow, the period between one shutting down and another starting up would increase. As flow increases beyond this level then overlap starts to occur. With flow levels of 228.957 mgd, then two centrifuges would be running almost continuously.

To generate data on the number of centrifuges one would take the amount wastewater that arrived at the plant 48 hours previous and divide by the mean flow level of 110.189 mgd. The number on the left side of the decimal point is the number of centrifuges that are always running, while the number on the right of the decimal, multiplied by 60, is the number of minutes during which they overlap. An example of this is shown below, using the flow on August 1 as input. Note that according to Equation 4.9 there are 0 centrifuges that are always running, and, according to Equation 4.10, there is 46 minutes of overlap. Since a centrifuge that is just being started cannot overlap one that is not running, this simply means that there is about 14 minutes of time in between centrifuge runs. Remember that this is a delayed signal, and this would be detected around midday on August 3rd.

$$\begin{aligned} Centrifuges &= 84.37/110.189 \\ Centrifuges &= 0.763 \\ NumberAlwaysRunning &= 0 \qquad (4.9) \\ Overlap &= 0.763 \cdot 60 = 45.76 \qquad (4.10) \end{aligned}$$

Sludge Trucks The sludge trucks are another form of delayed signal. According to the SME typically 8-12 Waste Management trucks pick up the sludge about 3

days after the wastewater enters the plant. Mapping that to the flow information, with 0 trucks coming with no flow, 8 trucks coming when flow levels are at the low end of medium flow mode (72.56 mgd), 12 trucks coming at the high end of medium flow mode (148.68 mgd), and 16 trucks coming during relatively high flows (186.74 mgd), the linear relationship shown in Equation 4.11 can be generated. Since this pickup is done over an entire day, it is necessary to compare the number of trucks to the average flow over the entire day 3 days earlier.

$$Trucks = round(0.0815 \cdot AverageFlow + 0.691) \tag{4.11}$$

Secondary Settling Tanks On one of the tours of the plant it was noticed that the first two secondary settling tanks were rather cloudy (see Figure 4.9). The SME said that the cloudiness in those tanks does tend to change from time to time. This is due to the delicate balance that needs to be maintained in the primary settling tanks and aeration tanks. When this balance is well maintained the secondary settling tanks are almost clear. If there is a slight misbalance then the southern most secondary settling tanks can get to be quite cloudy. This cloudiness is more common during periods of high flow mode (Lukas, 2007).

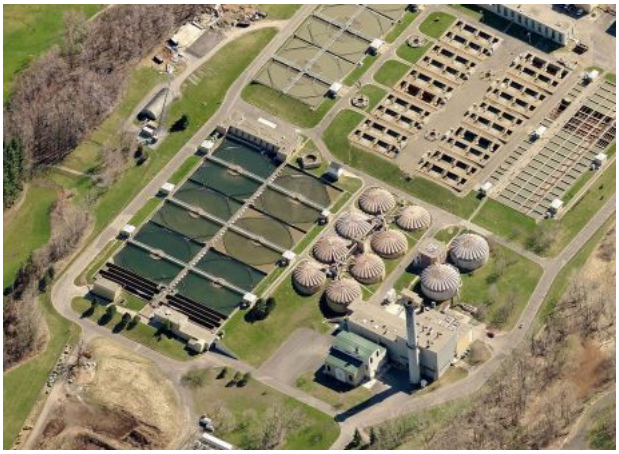


Figure 4.9: *An image of the secondary settling tanks on day when the biological balance was not maintained perfectly and they ended up very cloudy. Image courtesy of Bing Maps.*

One can simulate this in one of two ways. Increased cloudiness means less light is transmitting through the water and more is being reflected. That means the cloudiness could be a measure of reflectance. On the other hand, cloudiness also means that there is more dissolved organic material in the water. That means the cloudiness could be a measure of the approximate levels of CDOMs measured doing constituent retrieval analysis on multispectral images. Since reflectance is a simple variable requiring only panchromatic coverage and limited calibration that is being used here.

During periods of very low flow (50 mgd) the water is clear and the tanks appear black from up above, meaning there is almost 0 reflectance. In the event of epic failure at maintaining the delicate biological balance, the first secondary settling tank would have a reflectance that is almost identical to the last primary settling tank, which, using the empirical line method outlined in (Schott, 2007), averages around 24% for the visible region. Since, for most cases, it is assumed that failures are not present, that value will be halved and a reflectance of 12% is assigned to the highest flow level achieved in the flow data (430.81 mgd). One can then use the two points to form the relationship in Equation 4.12.

$$Reflectance = 0.0003 \cdot Flow - 0.016 \tag{4.12}$$

Alternative Generation This last group of variables were all generated based on flow information, which makes it awkward to use them to predict flow. They will instead have to be simulated based on the other forms of real data using the procedure outlined to generate tank information done earlier in this section. The algorithms look similar but each have some small differences due to the nature of each signal. The reflectance signal does not change based on rain and time but instead the probability of the reflectance being high changes, because the secondary settling tanks can be cloudy at any flow level, it is just more likely during periods of high flow. The thermal ratio from the transformers decreases during periods of rain to show that it cools down, and it warms back up after the rain has passed. The number of pumps running is normally fairly low, but has an increasing probability

of having a period of more pumps running over time and during periods of high flow. The centrifuges and trucks stay fairly constant with probabilities of increasing after days of rain and in the spring, and a probability of decreasing after periods of low precipitation and in the fall.

4.1.2 Testing with Real Data

The bulk of the tools use data from observations that were unavailable to us in any significant numbers. These will be treated shortly. However, one may want to see if the limited real data at least makes sense using the general approach. To do this it is best to start with the simplest case of predicting the flow mode using real data. There are three data collects with truth information: July 24, 2007, August 1, 2007, and May 3, 2008. The probabilities for all instances of each variable over the course of a year are known, but some of the variables are discrete and some are continuous. In order to use the conditional probabilities, the data will have to be sampled into bins, which will be explained shortly. Also, since real hourly information about the number of inactive tanks is not available, the real data regression analysis and geometric analysis will have to be done on just time and rain. The tank data can, however, be used when generating a template based on the real data and the SME information.

The Template Approach

Beginning with time, low flow mode is most likely to occur in the late summer to early fall, and from 3 a.m. - 8 a.m. in the morning. High flow mode is most likely to occur in the spring, and from 5 p.m. - 10 p.m. at night. The rest are more indicative of medium flow mode. With rain there are the two options of either simply using how much rain occurred during the 12 hours leading up to the prediction point, or using all of the information (1,2, 3, ... 48, 60, 72) and projecting it onto the principal component vectors. And lastly, with tanks one can again assume that 4-8 inactive tanks means medium flow mode. More than that means low flow mode, and less than that means high flow mode. These templates are shown below in Figure 4.10.

	Low	Medium	High
Time of Day	3-8	23-2, 9-16	17-22
Season	Fall	Winter/ Summer	Spring
Rain PC1	<0.3	0.3-4.0	>4.0
Rain PC2	<-0.65	-0.65-0.8	>0.8
Rain PC3	<-0.15	-0.15-0.52	>0.52
Inactive Tanks	>8	4-8	<4

Figure 4.10: *Templates made to determine the rate of flow based on all real data. PC means principal component.*

In order to get the range of values for the principal component vectors one has to take each vector one at a time and compare its values to how often the plant is in low, medium, or high flow mode. There will be significant overlap for low and medium and medium and high, but little overlap for low and high. The value that is picked such that less than that value is low flow mode and greater than that value is medium flow mode is the one that has equal probability of being in both modes. For example, rain PC1 has a 49.8% chance of being in low flow mode and a 49.9% chance of being in medium flow mode when it has a value of 0.3. As the value decreases, the probability of low flow mode increases. As the value increases, the probability of medium flow mode increases.

Now the data can be used to see how each piece is classified for each day. In Figure 4.11 one can see the plant at each of the three instances. What is interesting is that the July and August collect shows that one of the large primary settling tanks from Figure 3.9(b) is inactive. This is very interesting to capture because this is a fairly uncommon event. Something must have happened that required this tank to have some maintenance performed. The big tanks are the same as four of the other tanks and will be counted as such.

(a) July 24, 2007 (b) August 1, 2007 (c) May 3, 2008

Figure 4.11: *The VNIR band of each of the three real data collects. Notice the inactive large primary settling tank in the July and August images.*

Based on the results shown in Figure 4.12 the template says that the plant is running in medium flow mode with no uncertainty for both the July and August collects. Since these are simple templates, the error is not easily quantified. It would be best to put a maximum probability on templates of 95%. The value of 95 is not entirely arbitrary as it has been a statistical convention for a long time, but other than the fact that it is really close to 100% while still allowing for a little bit of error it has no concrete backing (Field, 2009). For large numbers of states, the probability maximum may have to change, because more states will often bring about more similarity between states. This can yield a situation where the maximum probability calculated is less than 50%, but is still significantly higher than the other prospects (Walvoord, 2008).

The results for May says that there is a 66.7% chance of medium flow mode, but a 16.7% chance of both high and low flow modes. The July and August collects are both a little under 90 mgd at the time of the collect, while the May collect is up at 108 mgd. All 3 of these are in the medium flow mode range, but the July and August collects are very close to the low flow range. The disadvantage of templates is that there is not a calculable amount of error in these analyses.

There are a couple different options here for applying weights. Currently, rain is weighted as three times more important as the other variables, because there are

	Jul 24, 2007	Aug 7, 2007	May 3, 2008
Time of Day	Medium	Medium	Medium
Season	Medium	Medium	High
Rain PC1	Medium	Medium	Medium
Rain PC2	Medium	Medium	Low
Rain PC3	Medium	Medium	Medium
Inactive Tanks	Medium	Medium	Medium

Figure 4.12: *The results of putting real data into the templates shown in Figure 4.10.*

three principal components. So, for example, one could weight the rain components based on the amount of variability each one explains in flow. PC1 would have a weight of 0.807, PC2 would be 0.131, and PC3 would be 0.029. This changes the probabilities to 3.3%, 71.5%, and 25.2% for low, medium, and high modes, respectively.

$$P(Mode) = Time + Season + RainPC1 + RainPC2 + RainPC3 + Tanks \quad (4.13)$$

$$P(Low) = (1 \cdot 0 + 1 \cdot 0 + 0.807 \cdot 0 + 0.131 \cdot 1 + 0.029 \cdot 0 + 1 \cdot 0)/3.967 \quad (4.14)$$

$$P(Medium) = (1 \cdot 1 + 1 \cdot 0 + 0.807 \cdot 1 + 0.131 \cdot 0 + 0.029 \cdot 1 + 1 \cdot 1)/3.967 \quad (4.15)$$

$$P(High) = (1 \cdot 0 + 1 \cdot 1 + 0.807 \cdot 0 + 0.131 \cdot 0 + 0.029 \cdot 0 + 1 \cdot 0)/3.967 \quad (4.16)$$

The tank ranges are based on SME information. With limited data collects one may want to put only a small initial confidence in this variable, perhaps 0.25. As the data increases the quality of the estimates provided by the SME may prove to be high, in which case the confidence may increase. On the contrary, if the data provided by the SME seems to regularly contradict the other data then the confidence level may decrease, and a new SME might be required.

Applying Dempster-Shafer to this template yields completely different results. The variables will have reliabilities applied that is based on how often each variable is in the suggested mode when given the range shown on the template. These reliabilities are shown in Table 4.6. Using those reliabilities on the three dates yields the results shown in Table 4.7. The results produce a probability of medium flow mode on May 3 that is slightly different than the standard calculation. It is stating that it is more likely for four variables to be correct with two incorrect than it is for one variable to be correct (i.e. one is high, one is low) while the other five are incorrect. This also opens up a small window in the probability of medium flow mode for the other two collects, acknowledging that it is not a guarantee. The results are quite promising when one considers that none of the variables has an exceptionally high reliability.

Table 4.6: *The reliabilities of the variables used in an application of Dempster-Shafer theory in determine the probability of each mode.*

Variable	Reliability
Time of Day	0.5207
Season	0.5671
Rain PC1	0.6058
Rain PC2	0.6179
Rain PC3	0.6240
Inactive Tanks	0.7294

Table 4.7: *Applying the reliabilities in Table 4.6 to the data in Figure 4.10 produces the results in this table. As is demonstrated here, it is more likely for four variables to be correct with two incorrect than it is for one to be correct with five incorrect.*

	P(Low)	P(Medium)	P(High)
Jul 24, 2007	2.9%	94.1%	2.9%
Aug 1, 2007	2.9%	94.1%	2.9%
May 3, 2008	15.6%	71.8%	12.6%

The LWIR signal from the transformers outside of the pump house can be taken into account. The results shown previously in Table 2.5 show that a high ratio of

transformer signal to surrounding area signal means they are drawing a lot of power, and a low ratio means they are off. For the July 24 data set, only one transformer is on, but its signal is also very weak. The August 1 data set has a stronger signal than July 24, but there is still only one transformer turned on. The May 3 data set has a strong signal, and both of the transformers are on. One can ultimately conclude then, using the template method, that the July 24 and August 1 collects were running in medium flow mode, with only a small amount of the wastewater coming from Irondequoit. The May 3 collect was probably (66.7%) running in medium flow mode, with a significant amount of wastewater coming from Irondequoit.

As is evidenced by the May 3 collect, there are times where some variables may contradict others. In many cases this is allowed because it is likely that there will be large amounts of overlap in the different states of a facility. Still, it is possible to have a variable indicate a state that is different from the other variables and have it be more important than the others. For example, if the time, season, and rain variables all indicate medium flow mode, but all of the settling and aeration tanks are inactive. When the tanks are not moving wastewater through then the plant is shut down, and special cases such as this need to be noted when making up the templates.

The real strength of this approach is that not all of the data is required to do this analysis. It can easily be scaled to the amount of data available at a particular point in time. This is incredibly significant because it also means that the templates are extensible, so any new signals determined to be significant to the analysis can simply be added in and accounted for. This differs greatly from the other two methods soon to be discussed. Instead of a basic understanding of the signal and the desired range of values assigned to each state, the more complex methods require large quantities of data for each signal that is used.

The Geometric Approach

If the tanks are not used in the analysis due to the sparseness of the data then some of the other options become available. Beginning with the geometric approach

one will first want to put the large list of all the variables together into a single data set. Next run the k-means algorithm discussed in Section 2.4.1, using three cluster centers. Sometimes using random cluster centers may cause this run more than once, since the randomness of the k-means algorithm does occasionally produce poor results. For example, a two dimensional projection of poor results is shown in Figure 4.13. Since it is a fast algorithm (less than 10 seconds on most computers) this should not be a big problem, and it is less likely to occur with more dimensions of data. This can also be avoided by initiating the algorithm with non-random cluster centers, based off of SME data. These are obtained in the same manner in which the templates are created. The results are shown in Table 4.8 using only the time of day, season of year, and rain principal component data.

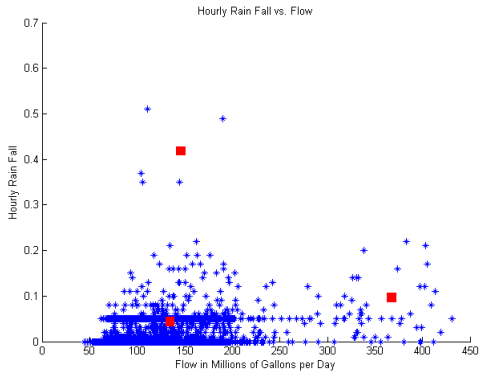


Figure 4.13: *A poor result from using random cluster centers to start the k-means algorithm. The cluster centers are shown as red boxes. 5 medium-high flow mode instances are placed in a cluster by themselves simply because they occurred during a period of high rain fall.*

Table 4.8: *Probability of Low, Medium, and High flow modes using the geometric approach on real data.*

Date	P(Low)	P(Medium)	P(High)
Jul 24, 2007	40.5	38.6	20.9
Aug 1, 2007	42.0	38.6	19.4
May 3, 2008	28.8	41.2	30.0

The Regression Approach

A regression model is rather simple to generate in this case, and using an ordinary least squares approach an R^2 value of 0.3469 was attained. The comparison of real values to modeled values are shown in Table 4.11 and are deceivingly accurate. Since time is a cyclical variable, it must be used differently than the other variables when performing a linear regression. One can find the mean of the flow for every hour of the day over the course of the year, and every month of the year, giving high and low frequency information about the flow. Each of these sets should then be demeaned, smoothed to reduce large jumps, (results shown in Tables 4.9 and 4.10) and subsequently added to the result of the regression equation.

Table 4.9: *Adjustments to the prediction of flow based on the month of the year. These are the values associated with m_i in Equation 4.18.*

January	17.098
February	21.616
March	16.659
April	13.865
May	3.983
June	-9.500
July	-17.640
August	-19.440
September	-17.960
October	-10.140
November	-4.039
December	5.499

Table 4.10: *Adjustments to the prediction of flow based on the hour of the day. These are the values associated with h_j in Equation 4.18.*

0	-0.058
1	-2.135
2	-4.527
3	-6.941
4	-8.989
5	-10.361
6	-10.635
7	-9.379
8	-7.101
9	-4.224
10	-0.911
11	2.116
12	4.432
13	5.839
14	6.677
15	7.124
16	7.149
17	6.637
18	6.199
19	5.657
20	4.883
21	3.950
22	2.943
23	1.653

Performing the regression on 9504 points using the rain information followed by making adjustments for time of day and year provided Equation 4.17, and was used to generate 9504 modeled points. The average distance the model data is from real data is 30.096 mgd.

$$Flow = 12.25 \cdot PC1 + 13.39 \cdot PC2 + 24.46 \cdot PC3 + 92.79 + Time \tag{4.17}$$

where

$$Time = m_i + h_j \tag{4.18}$$

The values for m_I and h_j are shown in Tables 4.9 and 4.10. Again, these are smoothed versions of the average flow over the course of a year for these time periods.

Table 4.11: *A comparison of real data to the values predicted by a regression model that used only real data. The numbers are in mgd.*

Date	Real	Modeled
Jul 24, 2007	89.89	90.23
Aug 1, 2007	84.37	91.42
May 3, 2008	108.93	109.63

The Conditional Probability Approach

As mentioned before, some of the variables are continuous and need to be sampled into bins in order to be able to use the conditional probability approach. Since having the rain be uniformly divided is not important, equal sized bins will be used because they are one step easier to set up. If one uses 6 bins for each of the rain principal components simply by taking the range of each one and dividing by 6, and then includes the 24 hours of the day and the 4 seasons of the year, this yields 20736 different possibilities. Clearly this is too many to demonstrate effectively here, so

to keep things simple just the 96 possibilities from using 24 hours in the day and 4 seasons will be shown.

In Tables 4.12, 4.13, and 4.14 are the distributions of each of the occurrences of each mode given the time of day and the season of the year. These are retrieved by counting the number of times the plant was in each flow mode at each hour during each season. The number of occurrences of each case is then divided by the total number of data points, 9504.

Table 4.12: *The percentage of occurrences of Low flow mode with the time of day and the season of the year.*

P(Low)	Winter	Spring	Summer	Fall
0	0	0.0001	0.0005	0.0004
1	0	0	0.0012	0.0014
2	0.0001	0	0.002	0.0026
3	0	0.0001	0.003	0.0035
4	0	0	0.0033	0.0046
5	0	0.0008	0.0033	0.0045
6	0	0.0014	0.0033	0.0048
7	0	0.0015	0.0031	0.0043
8	0.0001	0.0008	0.0019	0.003
9	0	0.0005	0.0014	0.0022
10	0	0.0001	0.0012	0.001
11	0	0.0001	0.0007	0.0004
12	0	0.0001	0	0.0004
13	0	0	0	0.0003
14	0	0	0	0.0001
15	0	0	0	0
16	0	0	0	0
17	0	0	0	0
18	0	0	0.0001	0
19	0	0	0	0
20	0	0	0	0
21	0	0	0	0
22	0	0.0001	0	0
23	0	0.0001	0	0

The values from each table are then divided by the sum of all values in that cell location. For example, the probability of low at 7 a.m. in the winter is divided by the sum of the probabilities of low, medium, and high at 7 a.m. in the winter. This produces the tables shown in Tables 4.15, 4.16, and 4.17.

Table 4.13: *The percentage of occurrences of Medium flow mode with the time of day and the season of the year.*

P(Medium)	Winter	Spring	Summer	Fall
0	0.0096	0.0113	0.0037	0.0109
1	0.0099	0.0115	0.0031	0.0101
2	0.0099	0.0115	0.0023	0.0091
3	0.0106	0.0117	0.0012	0.008
4	0.0109	0.012	0.0011	0.0068
5	0.0109	0.0109	0.0011	0.0069
6	0.0106	0.0106	0.0012	0.0069
7	0.0107	0.0102	0.0012	0.0072
8	0.0105	0.011	0.0024	0.0086
9	0.0101	0.0109	0.003	0.0092
10	0.0098	0.0111	0.0031	0.0106
11	0.0095	0.0111	0.0037	0.0113
12	0.0095	0.0114	0.0043	0.011
13	0.0096	0.0114	0.0043	0.0113
14	0.0095	0.0115	0.0042	0.011
15	0.0095	0.0111	0.0042	0.0114
16	0.0094	0.0115	0.0043	0.0113
17	0.0084	0.0113	0.0042	0.0111
18	0.0088	0.0111	0.0041	0.0113
19	0.0088	0.0111	0.0042	0.0113
20	0.0088	0.0111	0.0042	0.0114
21	0.009	0.0117	0.0042	0.0115
22	0.0094	0.0113	0.0042	0.0117
23	0.0094	0.0111	0.0043	0.0115

Table 4.14: *The percentage of occurrences of High flow mode with the time of day and the season of the year.*

P(High)	Winter	Spring	Summer	Fall
0	0.0027	0.001	0.0001	0.0012
1	0.0024	0.0008	0	0.0011
2	0.0023	0.0008	0	0.0008
3	0.0018	0.0005	0.0001	0.001
4	0.0015	0.0004	0	0.0011
5	0.0015	0.0007	0	0.0011
6	0.0018	0.0004	0	0.0008
7	0.0016	0.0007	0.0001	0.001
8	0.0018	0.0005	0.0001	0.001
9	0.0023	0.001	0.0001	0.0011
10	0.0026	0.0011	0.0001	0.001
11	0.0029	0.0011	0.0001	0.0008
12	0.0029	0.0008	0.0001	0.0011
13	0.0027	0.001	0.0001	0.001
14	0.0029	0.0008	0.0003	0.0014
15	0.0029	0.0012	0.0003	0.0011
16	0.003	0.0008	0.0001	0.0012
17	0.0039	0.0011	0.0003	0.0014
18	0.0035	0.0012	0.0003	0.0012
19	0.0035	0.0012	0.0003	0.0012
20	0.0035	0.0012	0.0003	0.0011
21	0.0034	0.0007	0.0003	0.001
22	0.003	0.001	0.0003	0.0008
23	0.003	0.0011	0.0001	0.001

Table 4.15: *The conditional probabilities of Low flow mode given the time of day and the season of the year.*

P(Low)	Winter	Spring	Summer	Fall
0	0	0.011	0.125	0.0326
1	0	0	0.2812	0.1087
2	0.011	0	0.4688	0.2065
3	0	0.011	0.6875	0.2826
4	0	0	0.75	0.3696
5	0	0.0659	0.75	0.3587
6	0	0.1099	0.7273	0.3804
7	0	0.1209	0.697	0.3478
8	0.011	0.0659	0.4242	0.2391
9	0	0.044	0.303	0.1739
10	0	0.011	0.2727	0.0761
11	0	0.011	0.1515	0.0326
12	0	0.011	0	0.0326
13	0	0	0	0.0217
14	0	0	0	0.0109
15	0	0	0	0
16	0	0	0	0
17	0	0	0	0
18	0	0	0.0303	0
19	0	0	0	0
20	0	0	0	0
21	0	0	0	0
22	0	0.011	0	0
23	0	0.011	0	0

Table 4.16: *The conditional probabilities of Medium flow mode given the time of day and the season of the year.*

P(Medium)	Winter	Spring	Summer	Fall
0	0.7802	0.9121	0.8438	0.8696
1	0.8022	0.9341	0.7188	0.8043
2	0.8022	0.9341	0.5312	0.7283
3	0.8571	0.9451	0.2812	0.6413
4	0.8791	0.967	0.25	0.5435
5	0.8791	0.8791	0.25	0.5543
6	0.8571	0.8571	0.2727	0.5543
7	0.8681	0.8242	0.2727	0.5761
8	0.8462	0.8901	0.5455	0.6848
9	0.8132	0.8791	0.6667	0.7391
10	0.7912	0.9011	0.697	0.8478
11	0.7692	0.9011	0.8182	0.9022
12	0.7692	0.9231	0.9697	0.8804
13	0.7802	0.9231	0.9697	0.9022
14	0.7692	0.9341	0.9394	0.8804
15	0.7692	0.9011	0.9394	0.913
16	0.7582	0.9341	0.9697	0.9022
17	0.6813	0.9121	0.9394	0.8913
18	0.7143	0.9011	0.9091	0.9022
19	0.7143	0.9011	0.9394	0.9022
20	0.7143	0.9011	0.9394	0.913
21	0.7253	0.9451	0.9394	0.9239
22	0.7582	0.9121	0.9394	0.9348
23	0.7582	0.9011	0.9697	0.9239

Table 4.17: *The conditional probabilities of High flow mode given the time of day and the season of the year.*

P(High)	Winter	Spring	Summer	Fall
0	0.2198	0.0769	0.0312	0.0978
1	0.1978	0.0659	0	0.087
2	0.1868	0.0659	0	0.0652
3	0.1429	0.044	0.0312	0.0761
4	0.1209	0.033	0	0.087
5	0.1209	0.0549	0	0.087
6	0.1429	0.033	0	0.0652
7	0.1319	0.0549	0.0303	0.0761
8	0.1429	0.044	0.0303	0.0761
9	0.1868	0.0769	0.0303	0.087
10	0.2088	0.0879	0.0303	0.0761
11	0.2308	0.0879	0.0303	0.0652
12	0.2308	0.0659	0.0303	0.087
13	0.2198	0.0769	0.0303	0.0761
14	0.2308	0.0659	0.0606	0.1087
15	0.2308	0.0989	0.0606	0.087
16	0.2418	0.0659	0.0303	0.0978
17	0.3187	0.0879	0.0606	0.1087
18	0.2857	0.0989	0.0606	0.0978
19	0.2857	0.0989	0.0606	0.0978
20	0.2857	0.0989	0.0606	0.087
21	0.2747	0.0549	0.0606	0.0761
22	0.2418	0.0769	0.0606	0.0652
23	0.2418	0.0879	0.0303	0.0761

4.1.3 Testing with Simulated Data

This section will begin with implementing the statistical and geometric approaches using the simulated tank data described in Section 4.1.1. When treating the simulated variable as equal with all of the other variables, neither method appeared to be adversely affected. In Table 4.18 the probabilities of each mode shifted slightly. In Table 4.19 it is clear that the May 3 model is now very far off. However, the average error in model data decreased to 26 mgd. This all happened because a large amount of real information went into making the simulated data.

Table 4.18: *Probability of Low, Medium, and High flow modes using the geometric approach with a simulated tank variable given equal weighting.*

Date	P(Low)	P(Medium)	P(High)
Jul 24, 2007	40.3	38.1	21.6
Aug 1, 2007	30.7	40.5	28.8
May 3, 2008	25.4	42.3	32.3

Table 4.19: *A comparison of real data to the values predicted by a regression model that used simulated values for a tank variable. The numbers are in mgd.*

Date	Real	Modeled
Jul 24, 2007	89.89	89.48
Aug 1, 2007	84.37	88.57
May 3, 2008	108.93	88.80

It is important to demonstrate the effect a poor simulation can have on the data. Instead of generating a logical set of values for the tanks, one could just generate a set of random integers from 0 to 20 that are not related to anything (a worst case scenario). Table 4.20 shows the difference between the results of giving the simulated variable equal weighting and instead giving it half the power of a real variable. As can be seen a poor simulation can simply wash out the differences between the classes, bringing about poor results, but the effects can be managed if the variable is not given equal weighting.

Table 4.20: *Probability of Low, Medium, and High flow modes using the geometric approach with a random variable given equal weighting on the left, and half weighting on the right.*

Weight	Equal			Half		
Date	P(Low)	P(Medium)	P(High)	P(Low)	P(Medium)	P(High)
Jul 24, 2007	37.5	36.1	26.5	38.4	37.5	24.0
Aug 1, 2007	37.2	35.9	26.9	42.0	38.6	19.5
May 3, 2008	31.2	37.7	31.1	29.7	40.3	30.0

The regression model is not as sorely affected as the geometric model. This is because the regression process applies weights to each variable. A random variable

that has nothing to offer to the model is given a very small weight. In fact, in doing ordinary least squares regression the random numbers were deemed insignificant, meaning they should not even be used in this model. Table 4.21 shows the results specific to the three days. As was expected with such a low weight being applied to the random variable, the modeled values return to about what they were using just the real data, and the R^2 value and error returned to their original quantities.

Table 4.21: *A comparison of real data to the values predicted by a regression model that used a random variable in the place of a simulated tank variable. The numbers are in mgd.*

Date	Real	Modeled
Jul 24, 2007	89.89	90.26
Aug 1, 2007	84.37	90.00
May 3, 2008	108.93	109.74

Since most of the data is in medium flow mode it is not too difficult to generate a method that will predict a point to be in that state. So, to help verify that these methods do work it is necessary to demonstrate how effective they are on simulated events of low and high modes (unfortunately no collects were ever acquired when the plant was in these modes). In Table 4.22 are five scenarios. Test 1 is pure low, test 2 is low with some conflicts, test 4 is high with some conflicts, test 5 is all high, and test 3 is highly conflicted data. All of these will be used in testing the predictive approaches.

The Template Approach

It is time to move on and get into a slightly more interesting series of tests using all of the simulated data combined with the real data. Take a look at the template shown in Table 4.23. When all of the variables are together like this it is clear that weighting becomes more of an issue. For instance, looking at the odor complaint row of the template, there should not be any complaints in both low and medium flow modes, and overall it is unlikely that one will show up during high flow mode. Basically, what this means is that if there is no complaint, this variable is worthless.

Table 4.22: *5 simulated scenarios that could happen at Van Lare. Test 1 is extremely low flow mode. Test 2 is low flow mode on the cusp of medium flow mode. Test 3 is conflicted data. Test 4 is high flow mode on the cusp of medium flow mode. Test 5 is extremely high flow mode.*

	Test 1	Test 2	Test 3	Test 4	Test 5
Time	5	12	5	12	20
Season	Fall	Fall	Spring	Spring	Spring
Rain PC1	0.1	0.5	0.1	3.8	4.5
Rain PC2	-1.0	-0.5	1.0	0.7	1.2
Rain PC3	-0.5	-0.4	-0.3	0.6	0.8
Inactive Tanks	12	10	2	4	0
LWIR Pumps	1.0	1.02	1.02	1.1	1.15
RF Pumps	0	1	4	3	4
Seismic Pumps	0	0	0	3	4
RF Centrifuges	0	1	2	2	3
Odor Complaint	0	1	0	0	1
Sludge Trucks	4	6	16	12	18
Reflectance	0	0.5	0	4.5	10

However, in the event of a complaint, this is a good indication of there having been high levels of flow. Because of this this variable should be ignored if there are no complaints, but given equal weighting to the other variables when there are complaints. The reflectance off of the secondary settling tanks will be treated in a similar manner.

The are three different variables associated with the pump house and trying to detect the amount of wastewater coming from Irondequoit. As was mentioned before, the number of pumps running is not a strong indication of operational mode unless there are a significant number of them running. This means that unless more than 4 pumps are detected as being active or that the LWIR signal around the transformers is exceedingly high, then without any further real data it is hard to justify giving these variables any weight. So, similar to the odor variable, when these signals are low or non existent, they will be ignored. When they are high, they will be given equal weight to other variables.

The sludge trucks variable is the number of trucks that came to the site to pick up sludge in a single day and is indicative of the average flow over the course of an entire day. Because of this that variable will get a weighting of 1/24, since the point in time where flow is being predicted is representative of a single hour. The centrifuges and the simulated tanks are pretty well understood so these variables will be given a weighting equivalent to how correlated each is to the flow, so 0.18 and 0.31, respectively. If real tank information is available then it will be given a full weight of 1.

Table 4.23: *A template representing all of the potential variables, real and simulated, being used to detect the flow level of the Van Lare plant.*

	Low	Medium	High
Time of Day	3-8	23-2, 9-16	17-22
Season	Fall	Winter/Summer	Spring
Rain PC1	<0.3	0.3-4.0	>4.0
Rain PC2	<-0.65	-0.65-0.8	>0.8
Rain PC3	<-0.15	-0.15-0.52	>0.52
Inactive Tanks	>8	4-8	<4
LWIR Pumps	<1.025	1.025-1.1	>1.1
RF Pumps	0-1	2-4	>4
Seismic Pumps	0-1	2-4	>4
RF Centrifuges	0-1	1	>1
Odor Complaint	None	None	1
Sludge Trucks	<8	8-12	>12
Secondary Settling Tank Reflectance	<1	1-4.4	>4.4

The results for the July 24, August 1, and May 3 data collects are present in Table 4.24 with real time, season, rain, tank, and LWIR data, while all other data is simulated. As can be seen here the signals do predominantly point towards medium flow mode for all three. Using the same weights on rain that were used before and ignoring the signals that are not giving any interesting results at the moment (odor, pumps, LWIR, reflectance) there are a total of 4.189 variables, 1 each for time, season, and tanks, 0.18 for centrifuges, and 0.042 for the sludge trucks. For July 24 and August 1 that yields P(Low)=1% and P(Medium)=99%. The May 3 data is slightly more exciting, with P(Low)=3%, P(Medium)=73%, and P(High)=24%.

Table 4.24: *A template of the results of combining real data with simulated data on the real data collects.*

	Jul 24, 2007	Aug 1, 2007	May 3, 2008
Time of Day	Medium	Medium	Medium
Season	Medium	Medium	High
Rain PC1	Medium	Medium	Medium
Rain PC2	Medium	Medium	Low
Rain PC3	Medium	Medium	Medium
Inactive Tanks	Medium	Medium	Medium
LWIR Pumps	Medium	Medium	Medium
RF Pumps	Low	Medium	Low
Seismic Pumps	Low	Medium	Low
RF Centrifuges	Medium	Medium	Medium
Odor Complaint	None	None	None
Sludge Trucks	Low	Low	Medium
Secondary Settling Reflectance	Low	Low	Low

Applying the Dempster-Shafer theory to this template using the reliabilities shown in Table 4.25 yields the results in Table 4.26. The results are very similar to the standard calculation method except for the July 24th data. This is due in

large part to the different weighting given to the variables. The standard calculation approach gives the simulated variables that are showing low flow mode almost no weighting, where the Dempster-Shafer approach has all variables on similar ground, varying slightly only in their reliabilities.

Table 4.25: *Reliabilities used for the Dempster-Shafer application shown in Table 4.26.*

	Reliability
Time of Day	0.5207
Time of Year	0.5671
Rain PC1	0.6058
Rain PC2	0.6179
Rain PC3	0.6240
Inactive Tanks	0.7294
LWIR	0.6702
RF Pumps	0.6076
Seismic Pumps	0.6076
Centrifuges	0.6237
Odor Complaints	0.8218
Sludge Trucks	0.6756
Reflectance	0.7684

Table 4.26: *Probability of Low, Medium, and High flow modes using the Dempster-Shafer approach on the simulated data.*

Date	P(Low)	P(Medium)	P(High)
Jul 24, 2007	23.4%	75.2%	1.4%
Aug 1, 2007	5.1%	94.2%	0.7%
May 3, 2008	19.3%	78.7%	2.0%

Using the five scenarios shown in Table 4.22 the template method holds up well. Table 4.27 shows the results of these tests. As can be seen this works really well with the pure cases (tests 1 and 5) and makes decent adjustments in the event of conflicting data. Test 2 is a low flow scenario but it has a an odor complaint present. The odor complaint shifts all of the probabilities in that direction, but low flow mode is still the most likely state of the plant. Test 4 is high flow mode with some of the

variables being at the high end of medium flow mode. This causes the expected shifts and splits the probability between high and medium flow modes. Test 3 is all conflicting data, and this is represented in the results: good probability of both low and high flow modes with no chance of medium flow mode. Such a scenario is only possible when there are data collection errors or something really strange is happening at the plant. Either way, the results clearly show that another look is needed in the event of such horrible data.

Table 4.27: *Probability of Low, Medium, and High flow modes for each of the simulated tests shown in Table 4.22 using the standard calculation approach of the template matching method.*

	Type	P(Low)	P(Medium)	P(High)
Test 1	Extreme Low	100%	0%	0%
Test 2	Low with Conflicts	45.9%	34.8%	19.3%
Test 3	Conflicting Data	35.4%	0%	64.6%
Test 4	High with Conflicts	0%	50.4%	49.6%
Test 5	Extreme High	0%	0%	100%

One could argue that the Dempster-Shafer approach gives better results in these scenarios than the standard calculation method. Looking first at Test 2 in Table 4.28 one can see that significantly less emphasis is placed on the fact that there was an odor complaint. Test 3 also shows some improved scoring, giving a higher probability to high flow mode.

Table 4.28: *Probability of Low, Medium, and High flow modes for each of the simulated tests shown in Table 4.22 using Dempster-Shafer theory.*

	Type	P(Low)	P(Medium)	P(High)
Test 1	Extreme Low	99.8%	0.1%	0.1%
Test 2	Low with Conflicts	95.6%	2.2%	2.2%
Test 3	Conflicting Data	47.9%	1.7%	50.5%
Test 4	High with Conflicts	1.6%	38.9%	59.5%
Test 5	Extreme High	0%	0%	100%

The Geometric Approach

Using the simulated data and plotting it in a high dimensional space yielded some poor results. The real data points (rain, tanks, LWIR signal) are normalized to a scale from 0 to 1. The simulated variables (pumps, reflectance, centrifuges), in order to be given less weight, are given a smaller range, .25 to .75. With a good description of low, medium, and high flow modes, a point for each mode is generated (and adjusted as per the normalization criteria). The Mahalanobis distance each data point is to each mode point is calculated, and the probabilities of each mode are scaled for each point individually. The distances and probabilities for the July 24, August 1, and May 3 data points are shown in Table 4.29. The probabilities of each mode are significantly closer together because each variable contributes to each mode, unlike the template matching scenario where each variable will only contribute information to one mode. In fact, a point that is exactly equal to the medium mode cluster center still has over 20% probability of the other two modes. This just means that the differences between the three modes are not great enough for this method to accurately predict which one the plant is in at the time of the collect.

Table 4.29: *The probabilities of each mode using the geometric method with simulated data.*

Date	Low Dist	P(Low)	Med Dist	P(Med)	High Dist	P(High)
Jul 24, 2007	6.584	32.1	5.123	36.1	6.711	31.8
Aug 1, 2007	6.401	32.3	5.109	35.9	6.585	31.8
May 3, 2008	6.621	31.8	5.211	35.7	6.352	32.5

Odor was a difficult variable to handle in this situation. With odor complaints being present acting as a strong indicator of high flow mode, but no complaints not being a strong indication of any mode made it hard to find a reliable manner in which to use this data. Initially the variable was treated like the other simulated data, and scaled from .25 to .75, and the high flow mode data point was given a .75 value in this dimension. This proved to be a mistake because high flow situations with no odor complaint were seen as having a decreased chance of high flow

mode. Scaling the variable down to a range of .4 to .6 helped this situation, but it significantly downplayed the situation when a complaint was present. It seemed to work best when the variable was ignored in cases when there was no complaint, and using the 0 to 1 scale in cases where there was a complaint. This is a bit awkward to implement, but it simply means that some dimensions are ignored and the Mahalanobis distance is calculated differently on a point by point basis. So each situationally significant variable will need a different covariance matrix for these calculations. This is an acceptable solution in this case because the distances are not being compared amongst points. Each point is only interested in comparing its distances to the three cluster centers to determine the probability of belonging to each cluster.

Applying this method to the five test cases yielded the results shown in Table 4.30. The best results showed up in Tests 1 and 4. Test 1 is pure low and yielded a relatively high probability of that mode. Test 4 is supposed to be a not too extreme version of high flow mode and the results show just that. Test 3, being a conflicting data test, should produce erroneous results, and it does. Test 2 had its probabilities shift heavily towards high flow mode because of the odor complaint. The data for Test 5 turned out to be too high for this approach, making it a rather large distance away from all of the clusters, thus smoothing out the probabilities.

Table 4.30: *Probability of Low, Medium, and High flow modes for each of the simulated tests shown in Table 4.22 using the geometric approach.*

	P(Low)	P(Medium)	P(High)
Test 1	42.4%	32.2%	25.5%
Test 2	32.7%	31.3%	36.0%
Test 3	31.9%	34.6%	33.6%
Test 4	23.3%	37.7%	39.1%
Test 5	30.2%	32.2%	38.7%

The Regression Approach

The variables for the geometric method were modeled as per the algorithms laid out in Section 4.1.1, and the same data was used for regression analysis. The regression equation is shown in Equation 4.19. This is different from Equation 4.17 in that there is not a time adjustment. This is because time is already considered when generating the simulated variables. Unfortunately, this model is slightly less accurate, achieving an R^2 value of 0.3095, with an average error of 31.57 mgd. This is significantly more realistic, however, because the previous model was dominated by the scalar term, which brought most measurements to the mean flow amount. The results were then adjusted from this based on the mean hourly and time of year flows present in the data - essentially using the data to predict itself - and practically providing an ideal scenario. Using information about the site to generate the simulated variables and then predicting the flow is one of the key elements of this thesis. Achieving marks so close to an ideal case shows the power of this analysis.

$$\begin{aligned} &34.55(Constant) \\ &+61.06 \cdot RainPC1 \\ &+40.83 \cdot RainPC2 \\ &+32.74 \cdot RainPC3 \\ &-30.08 \cdot Tanks \\ &+1.91 \cdot Reflectance \\ &-12.33 \cdot Centrifuges \\ &+6.90 \cdot Pumps \\ &-3.31 \cdot LWIRTransformers \\ &+57.50 \cdot SludgeTrucks \end{aligned} \tag{4.19}$$

The results shown in Table 4.31 show that the July 24 and Aug 1 flows were over estimated significantly, but they are within the range of the error in model. The May 3 data was significantly closer. When examining the results of running the regression on all of the data the model has the greatest error when trying to predict the extremely high flows. This makes sense as there are so few of them that the

model essentially treats them as outliers and makes little effort to accommodate for them.

Table 4.31: *The results of the regression using Equation 4.19. The numbers are in mgd.*

Date	Real	Modeled
Jul 24, 2007	89.89	109.51
Aug 1, 2007	84.37	102.34
May 3, 2008	108.93	117.98

Applying this equation to the test cases shown in Table 4.22 provides the results shown in Table 4.32. These results are outstanding given the nature of the scenarios. Tests 1 and 5 are supposed to be at the two extremes of the flow ranges. When creating the model these extremes are treated as outliers since they show up so rarely in the data. Thus the model cannot fully reach the extremes that the plant will occasionally exist in. This makes the values for Tests 1 and 5 far from extreme, but still within the expected mode. Test 2 is a low flow mode that is close to medium and had a predicted value at the low end of medium flow mode. Test 3 is conflicted data and had a predicted flow very close to the mean flow value. Test 4 is a high flow mode that is close to medium and had a predicted value at the high end of medium flow mode.

Table 4.32: *The predicted level of flow for each of the scenarios show in Table 4.22 using the regression approach. The values are in mgd.*

	Predicted Flow
Test 1	60.68
Test 2	77.17
Test 3	123.25
Test 4	137.94
Test 5	154.29

4.1.4 Summary

The analyses in this section are interesting because they are trying to identify the subtle differences that can occur in normal operational activities. These qualities can be what differentiates a site from normal operational mode and nefarious mode. Further exploration of the plant operations can be done by examining the three other real modes in which the Van Lare plant could be. Shutdown mode is a fairly easy mode to detect, single side mode is of moderate difficulty, while bypass mode is nearly impossible. Single side mode and bypass mode were never detected throughout this project, while shut down mode never occurred. This means that while these are completely real modes, the scenarios described in the following section are only hypothetical.

4.2 Single Side Mode

Single side mode can refer to one of four possible scenarios. The wastewater could be coming exclusively from the city or Irondequoit, or either of the two sides of operation could be handling all of the work. The wastewater coming exclusively from the city would be the easiest to detect. There would be no detectable activity coming from the pumps in the pump house - no RF, no seismic, no LWIR. Wastewater coming exclusively from Irondequoit would be slightly more difficult, but would reveal itself by having high amounts of activity in the pump house and would likely require a seismic sensor over the south side of the plant to make sure no flow was coming from that direction.

If only one side of the plant is running it becomes a slightly more interesting problem because all plant operations will still look fairly normal but only one set of settling tanks would be in operation and one of the grit removal buildings would also be out of commission. As demonstrated in Figure 4.11 it is fairly easy to tell when a settling tank is not being used when looking at it in the VNIR. To check to see if that side of the plant is not presently in use one would simply be able to check to see if all of the tanks are presently inactive. The other option would require

one to isolate the RF signatures of the machinery in each of the two grit removal buildings. Doing so would require the ability to get readings from both buildings up close when the equipment was on and off. This is an unlikely scenario unless it was obtained from onsite measurements.

Overall these binary cases, which may be important on occasion to know, do not make for very exciting tests. In an on/off case, no signal means its off, any signal means its on. Templates for these tests are trivial to make, as is running any sort of analysis.

4.3 Shutdown Mode

Shutdown mode is when there is no wastewater traveling through the plant. This does not mean, however, the plant has completely shutdown and no one is working there. In fact, some of the support activities could still be taking place, like pick-ups and deliveries, but the main functions of the plant would not be taking place. In the case of Van Lare, this means that there would be no active settling or aeration tanks, no active thermal signal on the transformers near the pump house or in the middle of the main facility, and no detectable RF signals other than radio communications.

Unlike mode detection, the RF signal is very important in this case. There are several pumps that can push the wastewater along through underground and in-building pathways directly out to lake Ontario. If a passive RF sensor is picking up any motor activity indicative of pump use then the plant is likely not in shutdown mode. If the plant is going through some major repairs and there are several pieces of construction equipment constantly in motion, then it might be difficult to determine whether or not detected signals are indicating pump use.

Since one is looking for the absence of signals, the presence of any signal is indicative of the plant not being shutdown. In such a black and white, binary scenario, a template test is a great method for detecting shutdown mode. Since one is looking for no activity, a template is fairly straightforward to design. Zero active tanks, 1:1 ratio of LWIR signal over the two transformer yards, and very

low RF detections, with the exception of the case where there is large amounts of construction activity. Such a template is shown in Table 4.33.

Table 4.33: *A basic template used to predict shutdown mode.*

	Shutdown
Active Tanks	0
LWIR Ratio: Pump house	1
LWIR Ratio: Main Site	1
RF Activity	Low
Construction Activity	-

The construction signal of the template is there to determine whether or not the RF signal is significant. For example, look at the 4 tests shown in Table 4.34. Test 1 is an ideal case where everything indicates shutdown mode and one could potentially say it is 95% likely to be in shutdown mode. Note: while a simple template may indicate 100% possibility of something it is rarely a good idea to ever say that something is 100% certain. In test 2 one can see that the RF activity is high, but so is the construction activity, meaning that detection of pumps is being interrupted by other signals. This essentially eliminates the variable as being useful, so one should just say that since three out of four of the variables indicate shutdown mode that it is 75% likely to be shutdown. Test 3 is a case where the RF is high even though there is no construction activity and none of the other variables are indicative of any activity within the plant. While only one variable states the plant is doing something, this is a case where one signal is all that is needed. This test indicates that the plant is 95% likely to NOT be shut down. The last test is simply a template of the plant on a normal day. This is a case where one can say that with 100% certainty that the plant is NOT shut down.

Table 4.34: *4 examples that use the shutdown mode template.*

	Test 1	Test 2	Test 3	Test 4
Active Tanks	0	0	0	8
LWIR Ratio: Pump house	1	1	1	1.1
LWIR Ratio: Main Site	1	1	1	1.1
RF Activity	Low	High	High	High
Construction Activity	None	High	None	None

4.4 Bypass Mode

Bypass mode is a rare occurrence where some of the wastewater is directed through grit removal and (probably) chemical treatment before heading out into the lake. The wastewater eligible for this is supposedly only from the storm drains from Rochester, so it should only be rain water (Monroe, 1998). Some of the conversations with plant employees indicated this was not entirely accurate, but it was never clearly stated to be untrue.

As mentioned in Section 3.1.3 bypass mode is something that takes place completely underground and is very difficult to detect. Significantly complicating the matter is that it only happens during significant rain events, during which time it is rather difficult to do remote sensing. In 1998 flows in excess of 135 million gallons per day (mgd) were allowed to be sent to the bypass, and flows exceeding 200 mgd were to be sent to an alternate treatment facility (Monroe, 1998). This is no longer the case, as the plant now has a capacity of 660 mgd - over three times larger than it was in 1998 (monroecounty.gov, 2011). If the bypass levels increased the same amount, then the bypass would be activated with flow levels greater than 445 mgd. This did not happen a single time in the hourly flow data obtained for this research that covers an entire year. Further, it was said that the bypass is very rarely used, sometimes going a year without being implemented (Lukas, 2007).

Essentially what that means is that if Rochester is experiencing a significant amount of flooding the bypass is on, otherwise it probably is not. Rochester has a huge underground storage system for wastewater, shown in Figure 3.1. This system can slowly build up wastewater for days and gradually send it to Van Lare for treatment. Irondequoit has a similar system as shown in Figure 3.2.

All of the underground pipes are shown in Figure 4.14. The labeled bypass pipes go through grit removal. One then throws its contents into the chemical treatment portion of the plant while the other goes directly out to Lake Ontario. Given that these are only used for bypass mode then there is not normally anything flowing through these pipes (Monroe, 2010). In the event that wastewater begins to flow through one would expect some slightly different vibrations to occur that might be detectable by a seismograph.

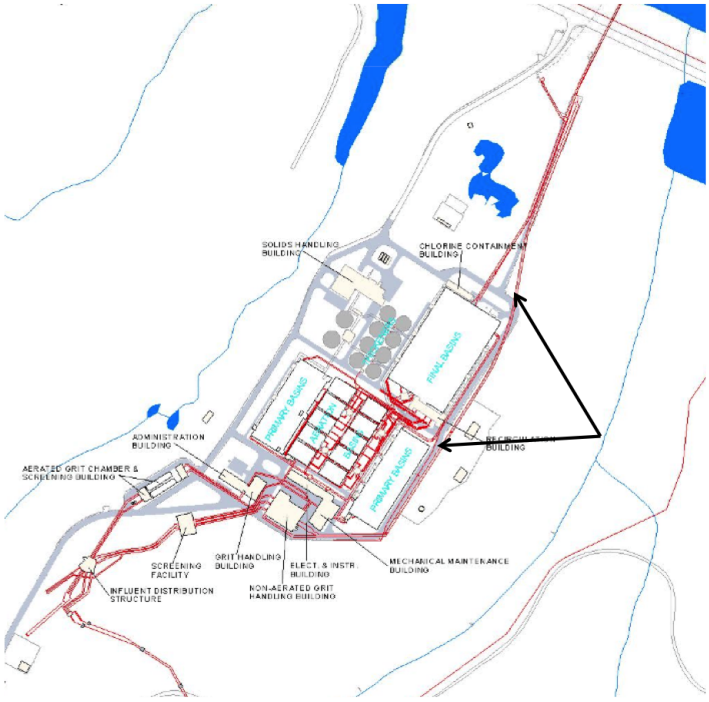


Figure 4.14: *An image of the underground pipes at Van Lare with the arrows pointing to the bypass pipes. The image is courtesy of Monroe County.*

The previous paragraphs cover the details of modes relevant to the Van Lare site.

To make this project more interesting I also explored some extreme modes that are not possible at Van Lare, but are generally possible at other industrial sites were explored.

4.5 Non-Likely Modes

What follows are some synthetic scenarios of the Van Lare site, with real analysis as to how an analyst should proceed if tasked with such a scenario at an alternative site. It is important to reiterate that the Van Lare facility is a well run wastewater treatment plant, and while some of the scenarios described here are possible, they are very unlikely and there is no evidence of any of them occurring.

4.5.1 Chemical Weapon Production

Chemical weapons are often in the form of dangerous gases that can be used to hurt or kill people without damaging property. They are more difficult to defend against than traditional weapons in that they affect an area of unknown dimensions, as wind will often be a factor in how they disperse. Knowing who has these weapons and tracking production is an important part of our nation's military defense.

Analyzing a site for chemical weapon production is not an easy process. Import, export, and production will be kept hidden, so subtle clues will have to be identified. The first step is to check the inputs that the plant already has. It is easier to hide weapon production by simply skimming a small percentage of the items a plant would normally receive and altering them. Van Lare receives regular amounts of sodium hypochlorite for disinfecting the wastewater. If it is combined with hydrochloric acid, a fairly common chemical, it creates chlorine gas and salt water. Chlorine gas was used as a weapon in World War I (Pulmonary Agents, 2011), and salt water is easily added to the wastewater on site.

Van Lare also receives regular amounts of activated carbon for the use of air sanitization. Chlorine gas, combined with the abundantly common carbon monoxide, can be passed through porous activated carbon to create phosgene, another

weaponized gas from World War I (Pulmonary Agents, 2011).

Detecting the production of such chemicals would require additional insights. An SME would have to have fairly intimate knowledge of how much sodium hypochlorite is needed based on the amount of flow. This means that the amount of flow will need to be known, with only a very small error, in order to prove beyond any doubt that the amount of sodium hypochlorite being delivered (a quantity that will also have to be known) is exceeding the amount the facility requires. The chemical reaction to make phosgene is exothermic, meaning that it releases heat and has to be cooled. It is typically done at a minimum temperature of 50°C, well above the highest temperatures seen in Rochester, NY (Pulmonary Agents, 2011). While a wastewater treatment facility has a very large amount of water accessible to use in the cooling process, the warmer wastewater will have to manifest itself somewhere. As discussed previously, all objects emit LWIR photons proportional to their temperature. Lastly, the spectral signatures of these gasses should be known and tested for, to see if there are increased abundances in any of the areas around the plant.

Suppose it is possible to get an accurate representation of the flow combined with sodium hypochlorite consumption, within a 5% error margin, and that the amount of hypochlorite being sent to the site is a known quantity. A chemical weapon template would look something like the one shown in Table 4.35. There are two known gases being searched for, as well as just a general template that is really a way of determining whether or not a site requires on site investigation.

In chlorine and phosgene detection the two key input materials are being monitored for excessive import based on the predicted being used. Sludge trucks, which come from another site, park inside a giant building on the Van Lare site (to be filled with sludge), and then drive off could easily be used to transport things on or off the site without being seen. If this was happening sludge truck traffic would exceed that of wastewater throughput. The amount of energy needed to power the site would be higher than anticipated if there were also several alternative processes taking place. Any leaks would lead to a small plume, or perhaps just an increase in the abundance of the chemical in the atmosphere. And, in the case of phosgene,

Table 4.35: *A basic template displaying some potential signals to look at in the pursuit of a chemical weapon investigation.*

	Chlorine	Phosgene	Unknown
Sodium Hypochlorite	Excessive	Excessive	Excessive
Sludge Pick-ups	Excessive	Excessive	Excessive
Electrical Usage	Excessive	Excessive	Excessive
Plume Detection	None	None	Type?
Thermal	None	Warm Water	Any
Activated Carbon	Normal	Excessive	Excessive

there needs to be some thermal signature in the wastewater on the plant.

Using Table 4.36 as an example, it is clear that while some signs point to the production of either chlorine or phosgene, there is an oddity that does not quite fit the model. The detection of a hydrogen plume was not discussed in the original assessment of the production methods of those two gasses. This does not mean necessarily that chlorine and phosgene are not being made, but it does seem to indicate that perhaps some other chemical is being produced, or that the production assessment was incorrect.

4.5.2 Environmental Hazard

Environmental hazards can manifest themselves in a number of ways, though sick or dead vegetation, sick or dead wildlife, and water contamination are three of the most common. Sick or dead vegetation usually means that either something has been leaked in the air or there is something in soil. Atmospheric problems can sometimes be identified with airborne hyperspectral imagery, tracking down the source of a gas leak. Slow leaks over years with small concentrations are rarely detectable and may require on-site measurements. Surface leaks in soil can also sometimes be detected

Table 4.36: *A basic template displaying collected signals to look at in the pursuit of a chemical weapon investigation.*

	Chlorine	Phosgene	Unknown
Sodium Hypochlorite	Yes	Yes	Yes
Sludge Pick-ups	No	No	No
Electrical Usage	Yes	Yes	Yes
Plume Detection	No	No	Hydrogen
Thermal	No	No	No
Activated Carbon	Yes	Yes	Yes

with hyperspectral imagery, but subsurface leaks would likely need soil samples for detection (Schott, 2007).

If a group of trees down wind of the Van Lare site gradually started to die, the different signals that would be worth knowing are soil pH, concentrations of gases and other elements in the soil and atmosphere, and a brief summary of the local wildlife inhabitants. A template for this scenario is available in Figure 4.15 which provides an example of the types of things an analyst would be looking for. The only type of data that could probably be detected remotely are the atmospheric constituents, the rest would probably require on site measurements. One would want to narrow down the possible causes of the tree deaths to be something in the atmosphere, in the ground, or biological.

The geometric test would likely work best in this scenario. Several types of soil contaminations, atmospheric contaminations, and biological contaminations are described in detail in various sources of literature. One can use these descriptions to simulate data for each of the signals and plot it in a high dimensional space, then measure the proximity of the real data point to several scenarios. If the real data is far from all of the different types of contaminations then the right one was not simulated. If the contamination appears to be in the soil, a leak is likely present and should be found. If it is in the air, it is possible that it comes from the site,

	Low	Ideal	High
Soil: pH			
Soil: Minerals			
Soil: Organic Materials			
Soil: Air			
Soil: Water			
Soil: CO2			
Soil: Nitrogen			
Oxygen			
Nitrogen			
Other Gases			
Vertebrates			
Invertebrates			

Figure 4.15: *A small collection of the items an analyst would want to investigate when trying to find the source of an environmental issue on or near a facility.*

but it is important to also check upwind of the site to make sure it is not from a different source. If the cause is biological, it may be a coincidence, or it might be that something about Van Lare makes it more possible for the organism to thrive. The process only tests to find out if, and it is not capable of explaining why.

4.5.3 Biological Hazard

This is a situation in which some chemical has leaked into the plant, either through the wastewater tunnel system or contact at the plant, and it has killed all of the biological content. This would cause the activated sludge process to fail completely and the plant would have to, somehow, empty its tanks and refill them with untainted wastewater to get the process going again. This is not a completely rare occurrence at smaller wastewater plants, but it would take huge concentrations of chemicals to have that big of an impact on the Van Lare plant. The aforementioned massive underground system has stations where testing is done before the wastewater gets to the plant, so bad chemical spills can be sent to the large holding areas where a large quantity of wastewater will join it (to dilute it) and make it no longer a problem. In the case of on plant contact, however, the first thing the plant would have to do is shutdown all of the influent and effluent pumps, trapping the wastewater within the tanks on the site. The tanks would then be shut down, and several trucks would have to come in to haul the tainted wastewater away so it can be treated at an alternate facility. Finally, a nearby facility would have to provide Van Lare with some mixed liquor, containing all of the biological material necessary to get the process started again (Bartlett, 2011).

In addition to massive truck traffic that is easily visible to both overhead and ground based sensors, seismic activity at the influent and effluent areas would cease. The reflectance of the wastewater would change, showing inactive tanks in the VNIR and hyperspectral sensors may be able to do constituent retrieval to find out which chemical caused the problem. Pumps and machinery would be turned off, bringing about little to no detectable RF signals. This would be short lived, however, as new RF signals would present themselves when pumping the wastewater out of all of the

tanks. This entire process would likely finish in less than a day, making it a very difficult, albeit interesting, phenomenon to detect.

4.6 Summary

This chapter has demonstrated the ability to perform process detection when merging multimodal data. Various analysis techniques can be used to join data from different modalities and provide a justifiable interpretation. Instead of saying "I am pretty sure X is happening" one can provide a probability of X, as well as the probabilities of the other possible states. This is very important for the intelligence community so that a defined level of action can hopefully, one day, be achieved for the various tasks set before it. It was also shown that there is a large margin of error present in these analyses, something that the government analysts seem to be leaving out of the press conferences demanding public support for military action. Much more work is needed to determine all the sources of this error and figuring out methods of minimizing it.

Chapter 5

Summary

This thesis outlines a process to determine the operational mode at an industrial facility. This method is designed to help take image exploitation a step further from target detection into process detection.

Three different methods for use in analyzing multi-modal variables have been examined. Overall it was shown that flow prediction is a very hard problem, and larger amounts of variability is needed in the different observable signals to yield better results. Template matching appeared to be the best at predicting the level of flow during normal operations, but again it is hard to show any error when doing that analysis. It was also shown to be the easiest test to implement, with strengths coming from its extensibility and allowing for missing data. A template could be set up in a matter of hours with the help of a subject matter expert and easily communicated to any analysts that are out in the field.

The geometric approach requires significant amounts of data, making it time consuming to set up. When a variable is not present at the time of a collect, a mean value can be substituted in to allow the data to be used, but it brings clusters closer together. Different clusters of data need to have little or no overlap in order for this method to have good results. Lastly, this method is not easily used in the field, making it difficult for real time analysis.

The regression approach is very powerful, and works well with simulated data,

but it requires large amounts of data from all possible plant scenarios in order to get an accurate model. This makes it extremely difficult to set up and only useful in long term surveillance scenarios. Once set up, however, the model and its accuracy are easily shared with the community, making it a powerful tool for real time analysis.

These methods are all data driven. The more data that is available the more accurate the results. Simulating data based on the input of an expert is helpful if done correctly, but can be detrimental if the simulation or information is poor. If data is not available it would be best to use a cooperative site to assist in building the models. With a few small adjustments made based on the differences between the cooperative site and the target site, data can be quickly ingested allowing for the quick development and testing of the analytical tools.

Lastly, this thesis provided multiple scenarios of the Van Lare plant and demonstrated how the different methods could be applied to get definite calculated probabilities instead of qualitative analytical feelings. The results of these analyses were fairly good, and were implemented with rather simplistic models of Van Lare operations. Imagine if all of the data RIT had collected over the past two decades of the Van Lare plant were able to be called up instantly. Instead of 3-6 data points, there would be nearly 100 data points. Perhaps even more variables would manifest and different modes would be even more distinguishable. The accuracy and confidence in the models would likely improve significantly.

This process, combined with the application of the analytical tools, 3D data registration, and an interactive computer environment, should improve future intelligence analysis. More work needs to be done developing ways to combine these tools in a way that is seamless and not cumbersome. Large difficulties still remain when it comes to bringing multi-modal data in to a single environment. Data entry is time consuming, but global data readers are also not feasible. Until that problem is solved the AANEE project will not truly be finished.

Chapter 6

Future Work

While the implementation in this thesis does demonstrate the utility of multimodal analysis, there are various projects that could be undertaken to improve the point.

6.1 Data Analysis

One way to build off of the work developed in this thesis is to use more sophisticated analysis techniques. Categorical regression is a form of multiple regression that incorporates nominal and ordinal variables. Some of the variables in this thesis fit that criteria, but were treated as continuous variables for simplicity.

Many of the relationships in this thesis between the variables and flow were assumed to be linear, when in reality they are not. Developing methods to get around this would be beneficial. Weighted regression is a form of multiple regression that is not just linear, but can be used for any functional relationship by providing additional weights to each of the variables (Statistics Methods, 2011).

At its basic form a neural network would be an iterative approach to come up with something resembling a regression equation. A more sophisticated version would develop a way to adjust the weights on the variables based on the value of each one, essentially developing a non-linear equation. A fuzzy neural network is simply doing this to data that has a wide range of potential values with the possibility of

overlap (Jones, 2008). This would make it an ideal item to pursue for this project.

6.2 A New Site

A major step in proving the utility of such analysis would be to run these tests at a different site. Not another wastewater treatment facility, but a different site entirely, with different functions, processes, and purposes. A nuclear power plant would be a great place to test and refine the methods developed in this thesis. Several parts of the process are again happening inside of buildings and are not easily detected. Multiple subject matter experts would likely be required to explain the various portions of the process that are used to generate and distribute electricity. Using the tools described in this thesis, would it be possible to determine the output of the facility? Is it possible, as the introduction of this thesis inquired, to determine whether or not a nuclear facility is doing something nefarious with a reasonable amount of error?

6.3 Data Over Time

A third study that would be useful to do, building from this thesis, is a detailed test of the impact of data over time. Using 4-5 different types of sensors tracking different signals, make collections continuously for a week/month/year. Sample this data at different levels and test how this affects the models. Test to see if the changes in the model affect mode detection, and if the way in which the data has an effect on the amount of error.

6.4 Building AANEE

Since this thesis is based on the idea that there is already a functioning AANEE environment, it would be prudent to lay out the plan for building such an environment. Building an environment in which to view data is easy enough, the real tasks

are in data ingestion and extraction, process model development and integration, and the development of a user interface that enables an analyst to easily use the various analytical tools mentioned in this thesis, as well as the countless others already available.

Data Ingestion and Extraction

Each data modality has its own typical manner in which it is stored in a file, with variations often present across different sensors or manufacturers. This leads to there being hundreds, if not thousands, of different file types for data ingestion into an AANEE environment. Modern programming techniques have mitigated the difficulty of creating so many readers. Each data type needs to be read in a specific manner, but the manner in which they are read remains fairly consistent: parse the header information and then read the data. This enables one to generate a generic reader function or class that is essentially a template for all forms of data. Since all data types already have some way in which to be read to a computer by whomever designed them, bringing the data into an AANEE environment becomes the lesser task of altering that data reader to fit the template. Such things have been implemented on a smaller scale already, in the ENVI environment for image data and the IViPP program using point data (Exelis, 2012; IViPP, 2012).

Data extraction is often done by way of letting a user select a region of interest, and then providing information about the data in that region. In the AANEE environment should a site have been heavily monitored for a few years, an analyst could potentially select a small region and still get mounds of information of various data types. Developers will need to work with analysts to determine what information is commonly desired and develop methods for easily bringing it out of the environment to the analyst in a format that is understood. This will be incredibly difficult, as each type of analyst is likely going to be most familiar with different types of data formats, and some advance query techniques will need to be implemented.

Process Models

Process models are simultaneously the largest and most challenging part of building an AANEE environment. Take, for example, the relationship among the electric transformers and the pump house on the Van Lare site. As more pumps turn on, more power is being drawn, so more transformers will need to be active and/or reach a higher temperature. Developing a physical model of this interaction will need to take into account the type of transformers and the type of motors powering the pumps. Different types of motors will require different amounts of power, and different types of transformers will respond uniquely when more power is demanded. This is just one tiny piece of the Van Lare puzzle and it will require electrical engineering experts with two different specialties just to create a simple model that can say, "If pump A turns on, transformer 1 does X." With all the various complex interactions taking place at sites of interest throughout the world, a vast number experts will be needed to build these models. It will also take a meticulous hand to make sure the models can interact with each other in an appropriate manner.

If all process models are in place for a site of interest then it becomes possible to run complex simulations based on a small amount of input. One can enter the state of a few variables at a given point in time and determine the possible range of signals for everything else at the site. This would provide a highly detailed and accurate account of the probability of each state of the site. This would be similar to the conditional probability approach previously discussed, with the massive amount of data required being replaced with complex physical models.

User Interaction

The main purpose of this environment is to make it easier for analysts to interpret the vast array of data presented to them. This will require a large array of various intelligence analysts to be included in almost every aspect of the development of the project. The tools and tricks they are used to having at their finger tips will need to be just as easy to do in a new environment or they will not use it. Similarly, the new tricks will need to be implemented in a way that is not confusing, which

can be difficult when one considers that many intelligence analysts are not technical people. Constant training and feedback will be required, making this a rather tedious process, but necessary in order to maximize the utility of AANEE.

Bibliography

[1] Allen, E. and J Iano (2008). *Fundamentals of Building Construction: Materials and Methods,* 5th Edition. John Wiley and Sons, New York, New York.

[2] *ArcGIS* (2010). Web. http://www.esri.com/software/arcgis/.

[3] Bartlett, V. (2007). Personal interview with the Head Trainer for the Environmental Training Center at Morrisville State College. Morrisville, New York.

[4] Bartlett, V. (2011) Personal interview with the Head Trainer for the Environmental Training Center at Morrisville State College. Morrisville, New York.

[5] *Bing Maps* (2011). Web. http://www.bing.com/maps.

[6] Bishop, Christopher M. (1999). "Latent Variable Models." *Learning In Graphical Models.* MIT Press.

[7] Borgatti, Stephen P. (1994). *How to Explain Heirarchical Clustering.* University of South Carolina.

http://www.analytictech.com/networks/hiclus.htm.

[8] Butt, Rizwan (2009). *Introduction to Numerical Analysis Using MATLAB.* Jones and Bartlett Learning.

[9] Commercial Real Estate Training Programs (2011). *Industrial Property Site Analysis.*

http://www.commercial-real-estate.net.au/industrial_property_site_analysis.html.

[10] Copper.org (2011). *Temperature Rise and Transformer Efficiency.* http://www.copper.org/applications/electrical/energy/trans_efficiency.html.

[11] *NIST/SEMATECH e-Handbook of Statistical Methods* (2011) http://itl.nist.gov/div898/handbook/index.htm.

[12] *ENVI Capabilities* (2012). Exelis Visual Information Solutions. http://www.exelisvis.com/ProductsServices/ENVI/Capabilities.aspx.

[13] *Design Considerations* (2011). Federal Pacific. http://www.federalpacific.com/university/transbasics/chapter5.html.

[14] Field, Andy. (2009). *Introduction to Statistics Using SPSS (and sex and drugs and Rock 'n' Roll).* Oxford University Press. Oxford, England.

[15] Gail, William B. (2007). "Remote Sensing in the Coming Decade: the Vision and the Reality". *Journal of Applied Remote Sensing* 1.1: 012505.

[16] "Google Earth" (2010). *Google.* http://earth.google.com/.

[17] "IViPP" (2012). SUNY Geneseo. http://cs.geneseo.edu/∼baldwin/ivipp/.

[18] Jing, Liping, Michael K. Ng, and Joshua Zhexue Huang (2007). "An Entropy Weighting k-Means Algorithm for Subspace Clustering of High-Dimensional Sparse Data." *IEEE Transactions on Knowledge and Data Engineering* 19.8: 1026-041.

[19] Jones, Edward R. (2008). *Neural Networks' Role in Predictive Analytics.* Information Management and SourceMedia. http://www.information-management.com/specialreports/2008_61/.

[20] Lukas, D. (2007). Personal interview with Senior Pollution Control Operator for Monroe County. Rochester, New York.

[21] Mehrotra, Rajiv. (1995) "Integrated Image Information Management: Research Issues." SPIE Digital Library. Visual Information Processing IV. Orlando, Florida.

[22] "Monroe County Wet Weather Operating Guidelines for the Frank E. Van Lare Wastewater Treatment Plant" (1998). New York State Department of Environmental Conservation. Rochester, New York.

[23] "Monroe County Wastewater: Collection and Treatment by Monroe County Operated Facilities" (2010). Monroe County Government.

http://www.monroecounty.gov/des-wastewater.php.

[24] "An Image Analysis and Software Development Environment" (2010). *NeatVision.*

http://neatvision.eeng.dcu.ie/index.html.

[25] O'Donnell, Erin (2005). *Detection and Identification of Effluent Gases Using Invariant Hyperspectral Algorithms.* Rochester Institute of Technology. Rochester, New York.

[26] *Open Clinical* (2010).

http://www.openclinical.org/home.html.

[27] Peebles, Peyton Z. (1980). *Probability, Random Variables, and Random Signal Principles.* New York, New York: McGraw-Hill.

[28] Petrie, Gordon and Gurcan Buyuksalih (2001). "Recent Developments in Airborne Infra-Red Imagers.." *Journal of Geo Informatics.*

[29] "Pulmonary Agents" (2001). *Federation of American Scientists.*

http://www.fas.org/nuke/guide/usa/doctrine/army/mmcch/PulmAgnt.htm.

[30] Rummel, R.J. (2010) "Understanding Correlation." University of Hawaii. http://www.mega.nu/ampp/rummel/uc.htm.

[31] Sentz, K. and S. Ferson (2002). *Combination of Evidence in Dempster-Shafer Theory.* SAND2002-0835 Technical Report. Sandia National Laboratories, Albuquerque, New Mexico.

[32] "The Industrial Facility Evaluation Program" (2011). *Schneider Electric.*

[33] Schott, John R. (2007). *Remote Sensing: The Image Chain Approach.* Oxford University Press. New York, New York.

[34] Secker, Jeff. (2005) "Exploitation of Multi-Temporal SAR and EO Satellite Imagery for Geospatial Intelligence." SPIE Proceedings.

[35] Sipser, Michael (1997). *Introduction to the Theory of Computation.* PWS Publishing. Boston, Massachusetts.

[36] Smith, F. Drew (2007). Personal interview with the Environmental Lab Technical Manager for Monroe County. Rochester, New York.

[37] "Open Source Image Analysis Environment" (2010). *TINA Vision.* http://www.tina-vision.net/.

[38] Walli, Karl (2003). *Multisensor image registration utilizing the LoG filter and FWT.* Rochester Institute of Technology. Rochester, New York.

[39] Walli, Karl (2010). *Relating Multimodal Imagery Data in 3D.* Rochester Institute of Technology. Rochester, New York.

[40] Walvoord, Derek J. (2008). *Advanced correlation-based character recognition applied to the Archimedes Palimpsest.* Rochester Institute of Technology. Rochester, New York.

[41] "Differential Diagnosis" (2011). *WebMD.*
http://dictionary.webmd.com/terms/differential-diagnosis.

[42] "Covariance" (2010). *Wolfram Mathworld.*
http://mathworld.wolfram.com/Covariance.html.